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

KASDI MERBAH UNIVERSITY OUARGLA
Faculty of hydrocarbons renewable energies and earth and universe sciences
Department of Earth and universe sciences

Master Dissertation
Specialization: PETROLEUM geology

Prepared by

- GUESSOUM Fedlellah
- TELLAB Saif eddin
- KOUIDRI Brahim

Theme

Enhanced Machine Learning Approaches for Predicting Compressional Sonic Logs in the Oglet En Nasser Field, Algeria

Evaluation Date: 11/06/2026

Jury Members:

Supervisor	HACINI Messoud	UKM Ouargla
Co-supervisor	BENFERDJALLAH Ayoub	UKM Ouargla
Jury Chair	LAOUINI Hamza	UKM Ouargla
Examiner	MEDJANI Fethi	UKM Ouargla

Academic year: 2025/2026

ACKNOWLEDGMENT

Alhamdulilla¯h Rabbil ‘A¯lam¯ın peace and blessings be upon Prophet Muhammad, his family, and companions. We thank Allah Almighty for His endless blessings, which enabled us to complete this thesis.

Sincere thanks to our beloved parents for their constant love, prayers, and encouragement, which have always been our greatest motivation. Our deepest gratitude is reserved for our esteemed supervisor, Professor Hacini Messaoud. His rigorous guidance and profound insights were the cornerstone of this research, shaping its direction and ensuring its scientific quality. We are immensely indebted to his mentorship throughout this thesis's journey. We also extend our sincere thanks to Mr. Benfarjallah Ayoub for his valuable assistance and continuous support. We are also pleased to express our sincere gratitude to the Head of the Geology Department for his wise guidance, continuous support throughout this work, and for providing us with the necessary data relevant to the substance of this study. Our thanks are also addressed to all our teachers for their academic generosity and the patience they have shown despite their heavy professional responsibilities. We extend our sincere appreciation to the esteemed members of the jury, Professor Medjani Fethi and Laouini Hamza, for their constructive feedback that will enrich and refine this research.

Finally, we extend our warm thanks to our colleagues and friends in the Geosciences Department and Sonatrach, as well as to our extended families, for their unwavering support. To everyone else who contributed, your kindness and encouragement are deeply appreciated.

ABSTRACT

Compressional sonic logs are indispensable for subsurface characterization, reservoir evaluation, and wellbore stability analysis, yet their acquisition is frequently constrained by operational limitations, high costs, and logistical challenges. This study addresses this gap by developing a machine learning framework to predict compressional sonic logs using conventional well log data from five wells. The methodology encompasses rigorous data preprocessing, including outlier removal, missing data imputation, and feature engineering to enhance predictive signal quality. Feature selection is performed to identify the most influential input variables, followed by training and validation of multiple regression algorithms: Random Forest, CatBoost, XGBoost, K-Nearest Neighbors, Support Vector Machines, and Deep Neural Networks. Hyperparameter tuning is systematically applied to optimize each model's performance, with evaluation conducted using correlation coefficients and root mean square error metrics. The ensemble-based models demonstrate the strongest predictive capability, achieving correlation coefficients between 83.6 and 88.4 percent and RMSE values ranging from 3.645 to 3.578. Additionally, the study confirms that proper input scaling is essential for distance-based and neural network models, while blind well testing further validates the generalizability and reliability of the predictions. This workflow offers a robust, cost-effective alternative for sonic log estimation in reservoir management and geomechanical studies

Keywords: Gamma Ray (GR) log, missing well log data, machine learning (ML), petro-physical parameters, reservoir evaluation, subsurface characterization.

RESUME

Les diagraphies soniques sont essentielles pour la caractérisation du sous-sol, l'évaluation des réservoirs et l'analyse de la stabilité des puits. Cependant, leur acquisition est souvent limitée par des contraintes opérationnelles, des coûts élevés et des difficultés logistiques. Cette étude vise à combler cette lacune en développant un cadre basé sur l'apprentissage automatique permettant de prédire les logs soniques compressifs à partir de données de diagraphies conventionnelles provenant de cinq puits.

La méthodologie adoptée comprend un prétraitement rigoureux des données, incluant la suppression des valeurs aberrantes, l'imputation des données manquantes et l'ingénierie des caractéristiques afin d'améliorer la qualité du signal prédictif. Une étape de sélection des variables a ensuite été appliquée afin d'identifier les paramètres d'entrée les plus influents, suivie de l'entraînement et de la validation de plusieurs algorithmes de régression, notamment Random Forest, CatBoost, XGBoost, et réseaux de neurones artificiels.

Un réglage systématique des hyperparamètres a été réalisé pour optimiser les performances de chaque modèle, tandis que l'évaluation s'est appuyée sur le coefficient de corrélation et l'erreur quadratique moyenne (RMSE). Les modèles basés sur les méthodes d'ensemble ont montré les meilleures performances prédictives, avec des coefficients de corrélation compris entre 83.6 % et 88.4 % et des valeurs de RMSE allant de 3.645 à 3.578.

Par ailleurs, l'étude confirme que la normalisation des entrées est essentielle pour les modèles basés sur les réseaux de neurones, tandis que le test sur un puits valide davantage la robustesse et la capacité de généralisation des modèles développés. Ce cadre de travail constitue une alternative robuste et économique pour l'estimation des logs soniques, avec des applications directes en gestion de réservoirs et en études géomécaniques.

مــــلخــص

تكتسب سجلات السرعة الانضغاطية الصوتية أهمية بالغة في تقييم المكامن النفطية وتحليل استقرار الآبار، إلا أن توفرها في جميع الآبار يظل محدوداً نظراً للتكاليف العالية والتعقيدات اللوجستية المصاحبة لعمليات التسجيل المباشر. ومعالجةً لهذه الفجوة، تقدم هذه الدراسة إطاراً حسابياً مبتكراً يستند إلى خوارزميات تعلم الآلة للتنبؤ الدقيق بهذه السجلات، مستفيدةً من البيانات المتاحة للمجسات التقليدية في خمسة آبار من آبار الدراسة. تتضمن المنهجية معالجة مسبقة شاملة للبيانات عبر إزالة القيم الشاذة، ومعالجة القيم المفقودة، وهندسة الخصائص لتحسين القوة التنبؤية للمدخلات. كما جرى تطبيق تقنيات اختيار الخصائص لتحديد المتغيرات الأكثر حساسية، ومن ثم تدريب ومقارنة أداء عدة نماذج وخضعت النماذج لعملية ضبط المعلمات الفائقة لتحسين كفاءتها، وتم تقييم أداءها باستخدام معامل الارتباط ، ومؤشرالخطأ وأثبتت النتائج أن نماذج التجميع هي الأكثر دقة في التنبؤ . وأظهرت النتائج أن تطبيع البيانات خطوة حاسمة لتحسين أداء النماذج المعتمدة على المسافات والشبكات العصبية، في حين أكد اختبار البئر الأعمى قدرة النماذج العالية على التعميم وموثوقيتها التدقيقية. ويقدم هذا الإطار في المجمل حلاً اقتصادياً وفعالاً لتقدير السجلات الصوتية الانضغاطية، مع تطبيقات عملية ملموسة في هندسة المكامن والتحليلات الجيوميكانيكية.

يمنح هذا المشروع شركة "سوناطراك" منظومة عمل قوية وواضحة تجمع بين المعرفة الجيولوجية والذكاء الاصطناعي، مما يفتح آفاقاً واسعة لتطبيقه الشامل في قطاع النفط والغاز.

Contents

CONTENTS

CHAPTER

METHODS AND MATERIALS

CONTENTS

CHAPTER4

RESULTS AND DISSCUTION

List of Figures

List of Tables

List of Abbreviations

GR: (Gamma Ray).

ML: (Machine Learning).

ANN: (Artificial Neural Network).

SVR: (Support Vector Regression).

RF: (Random Forest).

KNN: (K-Nearest Neighbors).

LR: (Linear Regression).

NPHI: (Neutron Porosity Log).

DTC: (Sonic Transit Time).

XGBoost :(Extreme Gradient Boost).

Phi: (Porosity).

Vsh: (Shale Volume).

Sw / SUWI: (Water saturation).

Sh: (Hydrocarbon saturation).

Rw: (Formation Water Resistivity).

K: (Permeability).

RHOB: (Bulk Density Log).

PE / PEF: (Photoelectric Factor).

CALI / CALX: (Caliper Log).

RXO: (Micro-resistivity).

K / POTA: (Potassium).

TH / THOR: (Thorium).

U / URAN: (Uranium).

M2R1 - M2R9: (Multi-depth resistivity).

TEMP: (Temperature).

MD: (Measured Depth).

TAGI: (Lower Triassic Clayey Sandstone Formation).

AEG: (El Gassi Clay).

Vp: (Compressional wave velocity).

MSE: (Mean Squared Error).

RMSE: (Root Mean Squared Error).

MAE: (Mean Absolute Error).

R: (Coefficient of Determination).

AI: (Artificial Intelligence).

SVM: (Support Vector Machine).

GUI: (Graphical User Interface).

gAPI: (Gamma Ray American Petroleum Institute unit)

RT / RT90: (Formation Resistivity).

EDA: (Exploratory Data Analysis).

General Introduction

Hydrocarbon reservoir exploration and evaluation constitute the cornerstone of the energy sector, relying fundamentally on accurate petrophysical and geophysical characterization of subsurface stratigraphic formations. Within this framework, well logs represent an indispensable tool for measuring the continuous physical response of penetrated formations, thereby enabling the identification of productive intervals and the assessment of the dynamic and mechanical properties of reservoir rocks. However, conventional deterministic methods used to estimate key rock-physics parameters such as porosity (ϕ), shale volume (Vsh), and water saturation (Sw) are generally based on simplified empirical equations (e.g., the Archie and Wyllie equations). These approaches often lose accuracy and general applicability when confronted with lithological heterogeneity and the structural complexities of unconventional reservoirs.

From a practical perspective, geophysicists frequently face the challenge of incomplete depth coverage in critical well logs due to poor borehole conditions (such as borehole collapse and washouts), mechanical failures of logging tools, or economically driven partial logging programs. Traditional practices commonly adopted to address these issues such as mean-value imputation or the exclusion of defective intervals particularly when dealing with historical datasets from Sonatrach, tend to distort cross-plots and weaken the statistical relationships among reservoir properties. This deficiency directly compromises the accuracy of subsequent geophysical modeling workflows, including acoustic impedance estimation and well-to-seismic tie analysis. Furthermore, the inconsistency of log mnemonics and the heterogeneous structure of legacy well files introduce additional challenges, requiring extensive data engineering, standardization, and quality control efforts before modeling can be undertaken.

General Introduction

In response to these limitations, contemporary geophysical research is increasingly shifting toward data-driven approaches. Machine learning algorithms offer the capability to capture the complex and nonlinear nature of rock-physics relationships, enabling the accurate reconstruction and generation of missing well logs while preserving their underlying physical consistency. Such approaches enhance the reliability of static reservoir modeling, petrophysical characterization, and geomechanical studies, ultimately improving decision-making in reservoir evaluation and development.

Our thesis is organized as follows:

• Chapter 1 presents a comprehensive analysis of the Oglet En Nasser reservoir, one of the major hydrocarbon accumulations in southeastern Algeria, highlighting its geological setting, reservoir characteristics.

• Chapter 2 provides an overview of Machine Learning, covering its definitions, fundamental concepts, various learning approaches (supervised, unsupervised, and reinforcement learning), and the principal methods and techniques employed in model development and prediction tasks.

• Chapter 3 summarizes well logging techniques used to estimate petrophysical properties in petroleum reservoirs. It covers both conventional methods and their limitations in complex geological settings. The chapter then highlights modern machine learning approaches, which improve the accuracy of compressional sonic log prediction and enhance reservoir characterization and decision-making in exploration and production.

• Chapter 4 concludes the study by applying machine learning methods and presenting both qualitative and quantitative results. It focuses on the prediction of compressional slowness (DT) using ensemble machine learning techniques and artificial neural networks (ANN), demonstrating their effectiveness in improving reservoir characterization and predictive performance.

• Finally, the General Conclusion synthesizes the key outcomes of the thesis, outlines their significance for Sonatrach and the oil and gas sector at large, addresses the limitations encountered during the study, and identifies potential directions for future research.

CHAPTER 1

GEOLOGICAL CONCEPT OF THE STUDY AREA

1 INTRODUCTION

The Oglet En Nasser field owes its high hydrocarbon prospectivity to a perfect synergy of tectonic and depositional histories. Structural traps formed during the Hercynian orogeny effectively seal and hold prolific hydrocarbons generated by deep Paleozoic source rocks within high-quality Triassic reservoir sands.

2 GEOLOGICAL CONTEXT

The Amguid-Messaoud Basin is one of the most prominent sedimentary basins within the Algerian Saharan Platform, covering a vast total area of approximately 157,793 km2. The basin is characterized by a diverse tectonic framework and a nearly complete stratigraphic succession, establishing it as a highly significant petroleum province known for its strategic reservoirs and a complex geological history that facilitated hydrocarbon maturation and accumulation [1].

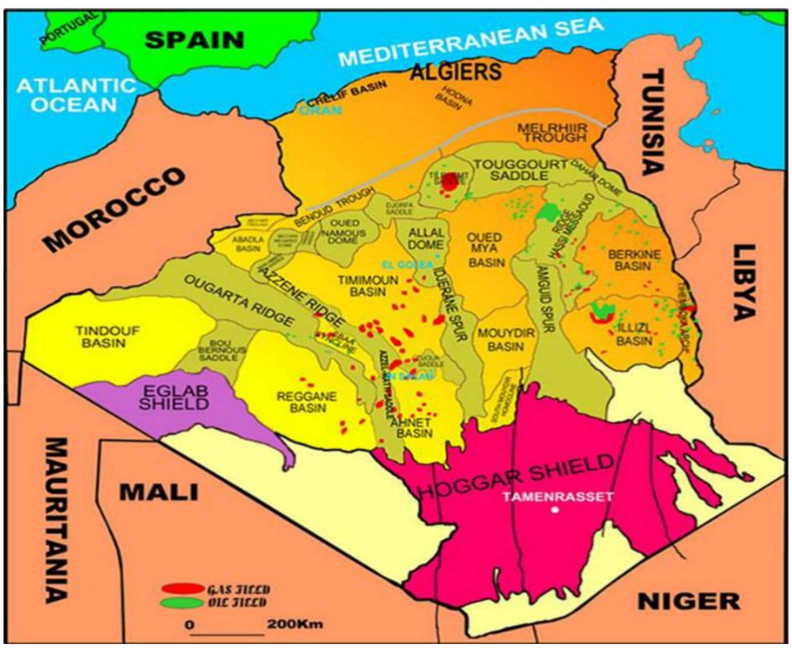

Figure 1. 1 : Sedimentary Basins of the Saharan Platform

3 GEOGRAPHICAL CONTEXT

The study area is situated within the Touggourt Est II perimeter, specifically encompassing operational blocks 415b and 424c of the Amguid-Messaoud Basin. Geographically, the region is located in the Lower Sahara (Southeastern Algeria), where the Oglet En Nasser structure emerges as a major geological feature-oriented Northeast-Southwest (NE-SW). The area is characterized by a flat desert landscape dominated by sand dunes and a hyper-arid Saharan climate.

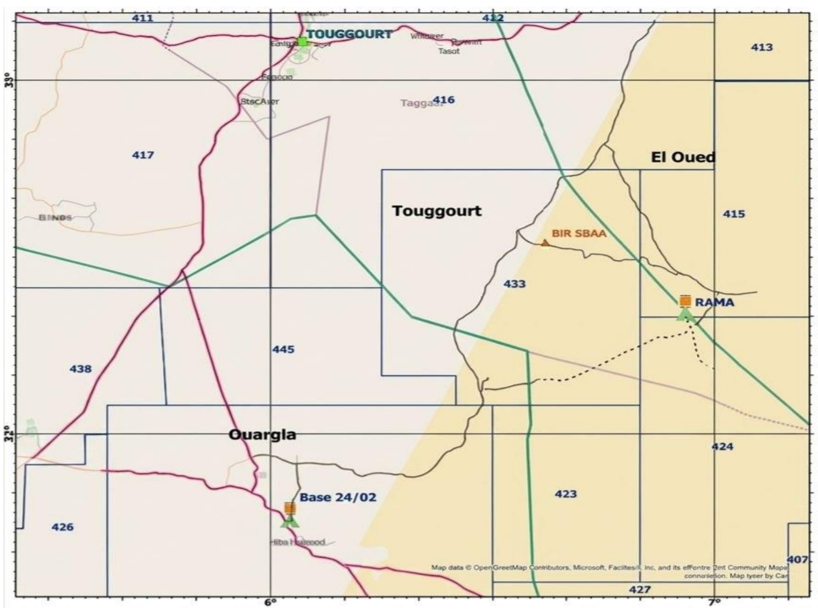

Figure 1. 2: Location map of the studied wells within Touggourt perimeter.

4 REGIONAL AND LOCAL GEOLOGY

The regional geology consists of a thick sedimentary cover ranging from the Paleozoic to the Quaternary, resting upon a Precambrian metamorphic basement. Geological evolution of this region was decisively shaped by the Hercynian orogeny, resulting in a major angular unconformity, followed by the deposition of Mesozoic and Cenozoic sequences, culminating in the superficial Mio-Pliocene deposits.

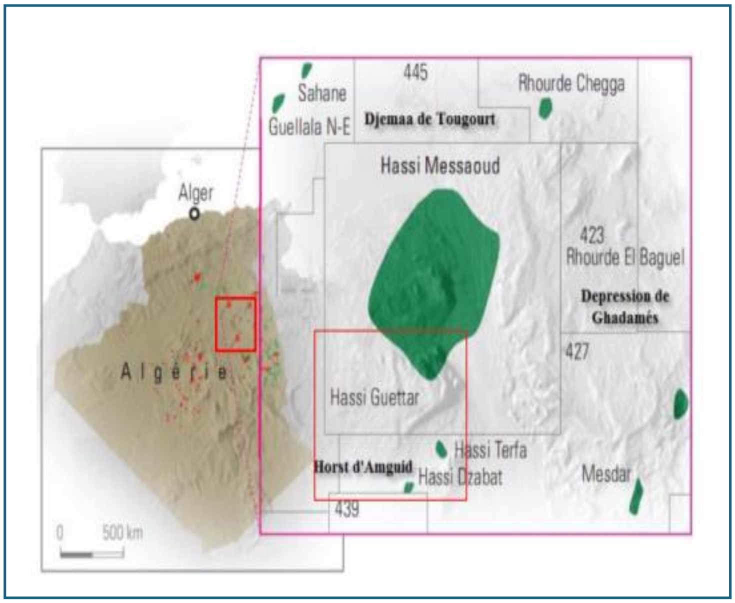

Figure 1. 3: Location map of the study area, Amguid [3].

5 STRUCTURAL FRAMEWORK

Structurally, the area is governed by a tectonic regime characterized by a series of Horsts and Grabens" delineated by major regional faults.The Oglet En Nasser structure is an elongated anticline covering an area of approximately 150 km 2with a structural amplitude of 50 meters.This configuration is influenced by a system of NE-SW trending normal faults with displacements ranging from 30 to 60 meters, complemented by E-W regional faults. This distribution reflects a clear tectonic control that significantly influenced the formation of structural traps and the distribution of reservoirs in the region [4].

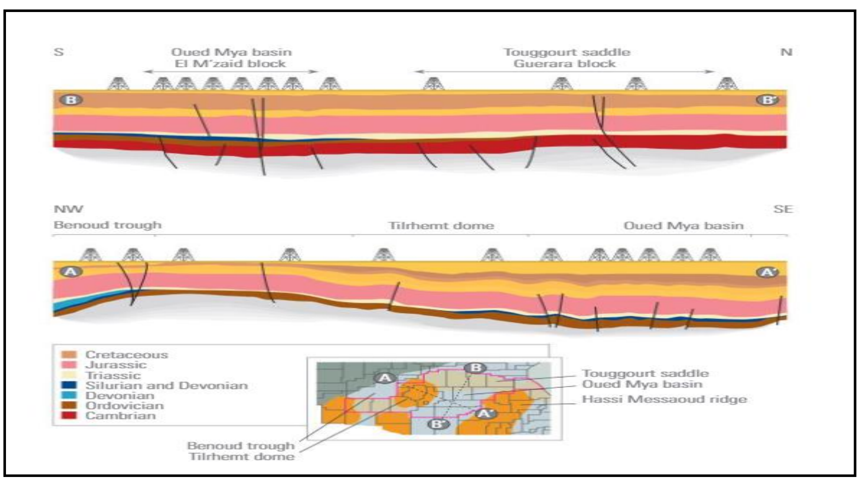

Figure 1. 4: The N-S and NW-SE sections in the Oued Mya basin [5].

6 LOCAL STRATIGRAPHY AND LITHOLOGICAL DESCRIPTION

The sedimentary cover in the Oglet En Nasser area, within the Amguid-Messaoud Basin, consists of a thick and diverse sequence ranging from the Cenozoic to the Paleozoic basement. The stratigraphic succession is detailed as follows:

❖ **Cenozoic Stratigraphy**

- **Mio-Pliocene:** Consists predominantly of continental deposits characterized by white to translucent, locally yellowish sands, ranging from fine to coarse-grained, with thin intercalations of friable carbonate sandstone.
- **Eocene:** Represents a transitional marine facies characterized by alternating white anhydrite and chalky limestone, with occasional occurrences of flint (silex) and white dolomite.

❖ **Mesozoic Stratigraphy**

A. Cretaceous

Carbonate Senonian: Consists of alternating dolomitic limestone and calcareous dolomite, with thin interbeds of marl and clay. The basal interval commonly includes white limestone and anhydrite with minor traces of gypsum.

Anhydritic Senonian: This unit is dominated by massive white anhydrite, interbedded with clay and argillaceous dolomite, reflecting deposition under strongly evaporitic conditions.

Saliferous Senonian: Comprises thick sequences of rock salt (halite) and anhydrite, with minor intercalations of gray-green silty clay and dolomite, indicative of highly restricted and hypersaline environments.

Turonian: Consists mainly of white to beige limestone, locally chalky or argillaceous, interlayered with thin gray-green clay partings, suggesting a relatively open marine setting with intermittent siliciclastic input.

Cenomanian: Characterized by an alternation of soft gray-green clays and white microcrystalline anhydrite, with meter-scale intervals of dolomitic limestone, reflecting fluctuating depositional conditions between clastic and evaporitic influences.

Albian: This unit is predominantly composed of fine- to medium-grained argillaceous sandstones interbedded with plastic clays, grading downward into coarser sand facies at the base. It represents a major regional aquifer and reflects deposition within a high-energy fluvial to deltaic system.

Aptian: Consists of white dolomitic limestone of moderate hardness, locally transitioning into beige microcrystalline dolomite, indicative of a shallow marine carbonate platform setting.

Barremian: Characterized by white to brownish-red coarse-grained sandstones, intercalated with brown silty clays and subordinate gray limestone layers, reflecting alternating continental to marginal marine conditions.

Neocomian: Comprises gray to reddish-brown soft clays interbedded with beige to gray-white sandstones, with occasional lignite traces, suggesting deposition in continental to deltaic environments with local organic-rich intervals.

B. Jurassic

Malm: Represents a heterogeneous succession of fine- to medium-grained siliceous sandstones, brown silty clays, dolomite, and anhydrite, reflecting variable depositional conditions ranging from continental influence to restricted marine settings.

Dogger: The Dogger interval is subdivided into two main facies: the Argillaceous Dogger, which is composed of variegated clays and glauconitic sandstones deposited in a shallow marine environment with moderate energy conditions, and the Lagoonal Dogger, which is dominated by massive anhydrite with interbedded dolomite and marl, reflecting deposition in a restricted evaporitic lagoonal setting.
Lias: Represents the main evaporitic sequence and is subdivided into several key units:

Anhydritic and Saliferous Lias: Alternation of pulverulent anhydrite, massive translucent halite, and plastic brown clay, indicative of hypersaline restricted basins.

Horizon B (LD3): A regionally recognized marker bed composed of gray argillaceous limestone.

Lias S1, S2, and S3: Thick successions of translucent to pinkish massive halite interbedded with anhydrite and red-brown clay partings.

Argillaceous Lias (G10): Red-brown saliferous clays containing fine halite inclusions, representing distal evaporitic deposits

C. Triassic

Trias S4 (G20): Dominated by massive translucent halite deposits with intercalations of indurated red-brown to chocolate-brown clays, reflecting deposition in highly evaporitic and restricted basin conditions.

Argillaceous Trias (G30): Comprises red-brown silty clays, locally dolomitic, interbedded with hard beige microcrystalline dolomite layers, indicating episodic carbonate precipitation within a predominantly clastic setting.

Trias T2: Characterized by an alternation of white to gray crystalline dolomite and red-brown silty clays. This unit is frequently associated with volcanic (eruptive) material, suggesting syn-depositional tectonic activity.

Trias T1: Represents an argillo-micaceous succession composed of fine-grained sandstones, siltstones, and laminated micaceous clays, indicative of low-energy continental depositional environments.

Eruptive Trias: Defined by volcanic deposits interbedded with indurated silty clays and dolomite, reflecting tectonically active conditions during sedimentation.

Lower Triassic Series: Constitutes the principal hydrocarbon reservoir in the study area, consisting of high-quality fluvial to deltaic sandstones characterized by excellent porosity and permeability, indicative of favorable reservoir properties.

❖ Paleozoic Stratigraphy

The Paleozoic succession unconformably underlies the Mesozoic sequence below the Hercynian unconformity:

Silurian: Dominated by dark, organic-rich shales representing the main source rock in the region. This interval often corresponds to the deepest drilled sections in several wells.

Ordovician: Composed of hard quartzitic sandstones and micro-conglomeratic clays, forming a secondary reservoir target with moderate storage and flow capacity.

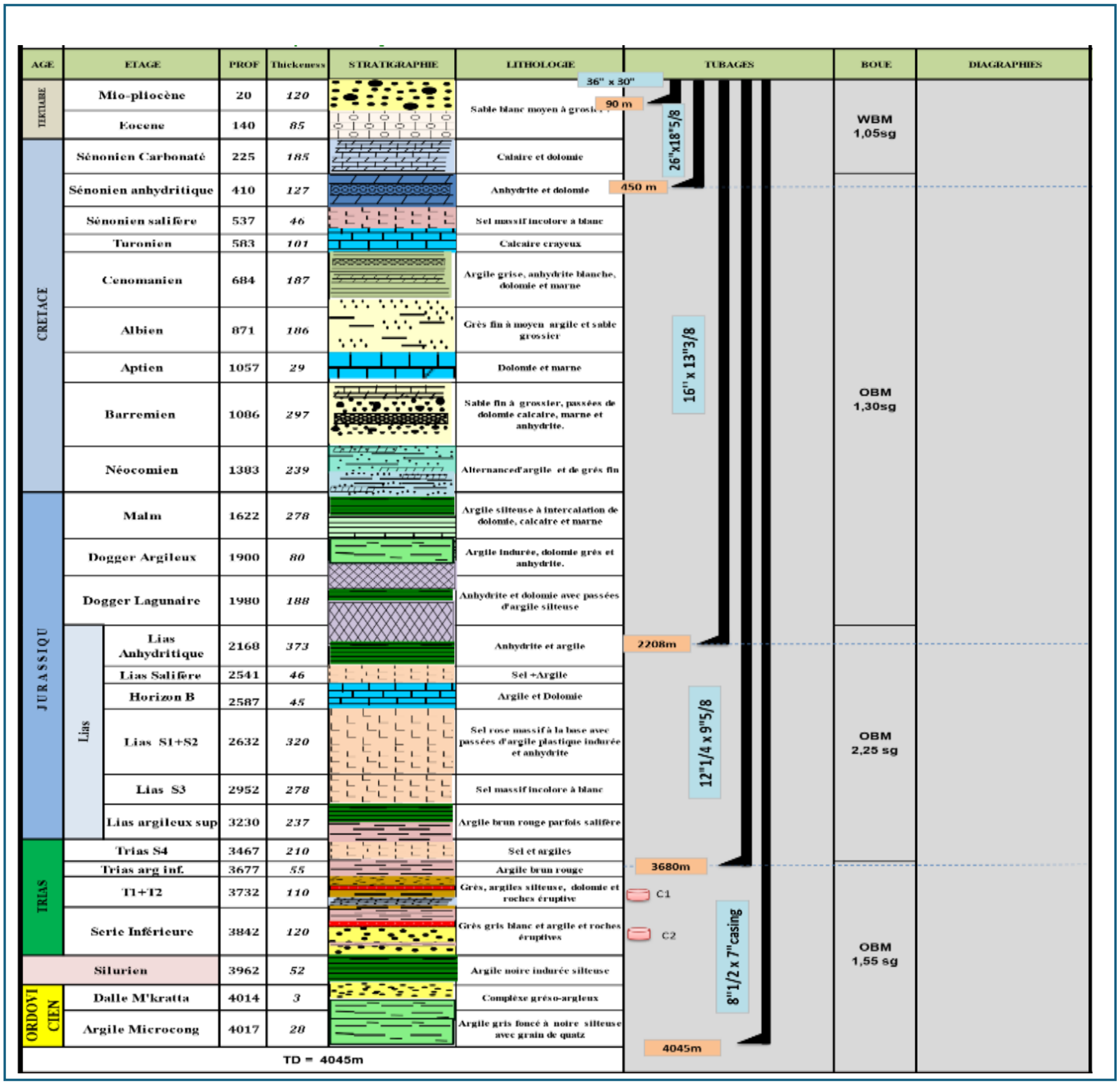

Figure 1. 5: Column field Stratigraphy of Oglet En Nasser area, within the Amguid-Messaoud

7 PETROLEUM SYSTEM OF THE OGLET EN NASSER

The petroleum system in the Oglet En Nasser area, situated within the Amguid-Messaoud Basin, is characterized by a high-performance synergy between structural evolution and stratigraphic sequences. The following components define this system:

Source Rocks

The primary source rock is the Lower Silurian (Llandoverian). These organic-rich shales possess high generation potential and have reached optimal thermal maturity within the Amguid-Messaoud depression, enabling the generation and expulsion of significant quantities of liquid and gaseous hydrocarbons.

Reservoir Rocks

Hydrocarbon accumulation occurs mainly in two stratigraphic levels: the primary reservoir is the Lower Triassic Series, composed of fluvial to deltaic sandstones with excellent porosity and permeability, as confirmed by productive wells, while secondary reservoirs include Ordovician sandstones and selected Triassic T1–T2 sandy intervals, which provide additional but comparatively lower permeability potential.

Seal Rocks (Cap Rocks)

The hydrocarbon system is effectively sealed by two main lithological barriers: an immediate seal formed by the massive Triassic evaporites and anhydrite layers, which act as a highly impermeable barrier directly above the Triassic reservoirs, and a regional seal provided by the Saliferous Senonian sequence, which ensures long-term containment and preservation of hydrocarbons within the system.

Trapping Mechanism

The traps in the area are mainly structural, formed by Hercynian tectonic activity, and consist of fault-bounded horsts and anticlinal closures. The Hercynian unconformity strongly controls reservoir distribution by juxtaposing reservoir sands against source or sealing formations, enhancing hydrocarbon trapping efficiency.

8 Conclusion

The Oglet En Nasser field is a major hydrocarbon province with a thick sedimentary record from the Paleozoic to the Cenozoic. Its geological framework was mainly shaped by the Hercynian tectonic phase, which created a complex system of faults, horsts, grabens, and anticlines. The stratigraphy includes siliciclastic, carbonate, and evaporitic units deposited in continental, marine, and restricted basin settings. This structural and depositional complexity strongly controls reservoir distribution and trap formation, making the field highly prospective within the Algerian Saharan Platform.

CHAPTER 2

FOUNDATIONS OF ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

1 INTRODUCTION

Artificial Intelligence (AI) and Machine Learning (ML) have become essential technologies in modern scientific and industrial applications. AI involves systems designed to perform tasks that require human intelligence, while ML enables systems to learn from data and improve performance without explicit programming. These technologies have significantly enhanced data analysis, pattern recognition, and predictive modeling. They are widely applied in various sectors, including healthcare, finance, energy, and transportation. In geosciences, AI and ML are used to process and interpret complex subsurface data. In the petroleum industry, they play a key role in seismic interpretation, well log analysis, reservoir characterization, and hydrocarbon exploration [6].These approaches improve the accuracy of petrophysical predictions and reduce uncertainty in geological models. They also enhance operational efficiency and support data-driven decision-making. Additionally, AI and ML help optimize exploration and production processes while reducing costs and risks. Their integration leads to more reliable and robust geological and geophysical interpretations. This chapter presents the fundamental concepts and evolution of AI and ML. It also highlights their growing importance in geophysics and petroleum engineering applications [7] [8].

2 SUMMARY OF ARTIFICIAL INTELLIGENCE

Artificial Intelligence (AI) is a field within computer science that focuses on developing intelligent systems capable of performing tasks that normally require human cognition, such as learning, reasoning, problem-solving, perception, and language understanding. The main goal of AI is to build computational models that can analyze and interpret data, make decisions, and adapt to changing environments with minimal human intervention [9]. AI systems are commonly classified into narrow AI, which is designed for specific applications, and general AI, which aims to replicate broad human cognitive abilities [8]. In modern applications, AI is closely integrated with machine learning techniques that enable systems to improve their performance through experience and continuous exposure to data. As a result, AI plays a significant role in automating complex processes, extracting meaningful patterns from large datasets, and supporting decision-making across various domains such as science, engineering, and industry [10][11].

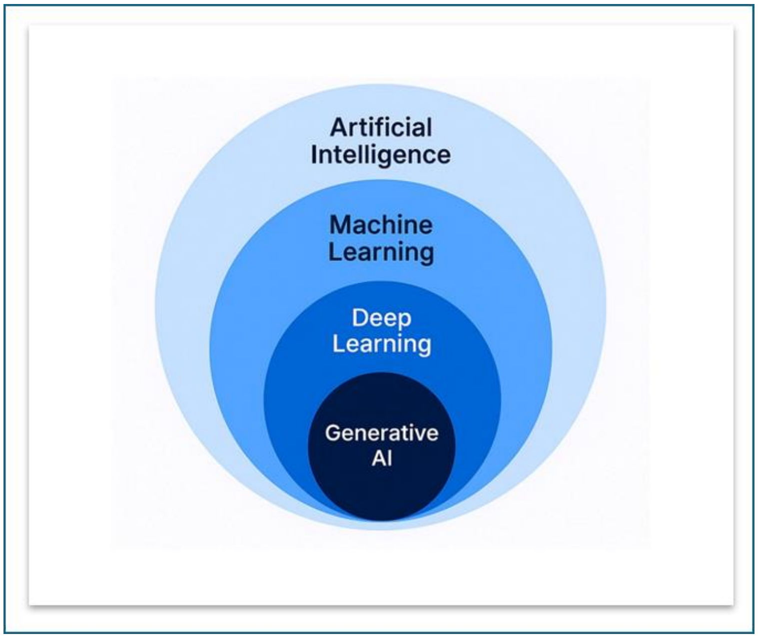

Figure 2. 1: Generative AI, Deep, Machine Learning and Artificial Intelligence relationship.

3 MACHINE LEARNING

The concept of Artificial Intelligence (AI) originated from early theoretical contributions such as Alan Turing's "learning machine" (1937) and the "threshold logic" model proposed by Warren McCulloch and Walter Pitts in 1943, which simulated the functioning of biological nervous systems. Among the most influential branches of AI is Machine Learning (ML), a field that has significantly impacted numerous scientific and academic domains while continuing to expand into new areas. Machine learning refers to algorithms capable of learning from data in order to accomplish specific tasks such as prediction or classification. Generally, ML systems operate through three main stages: (1) receiving and processing input data, (2) analyzing the data to identify underlying patterns and governing relationships, and (3) generalizing these learned rules to make accurate predictions or decisions on new, unseen datasets [12].

ML encompasses techniques such as supervised learning, unsupervised learning, and reinforcement learning, and it is widely applied in natural language processing, computer vision, robotics, and healthcare [10] [13].

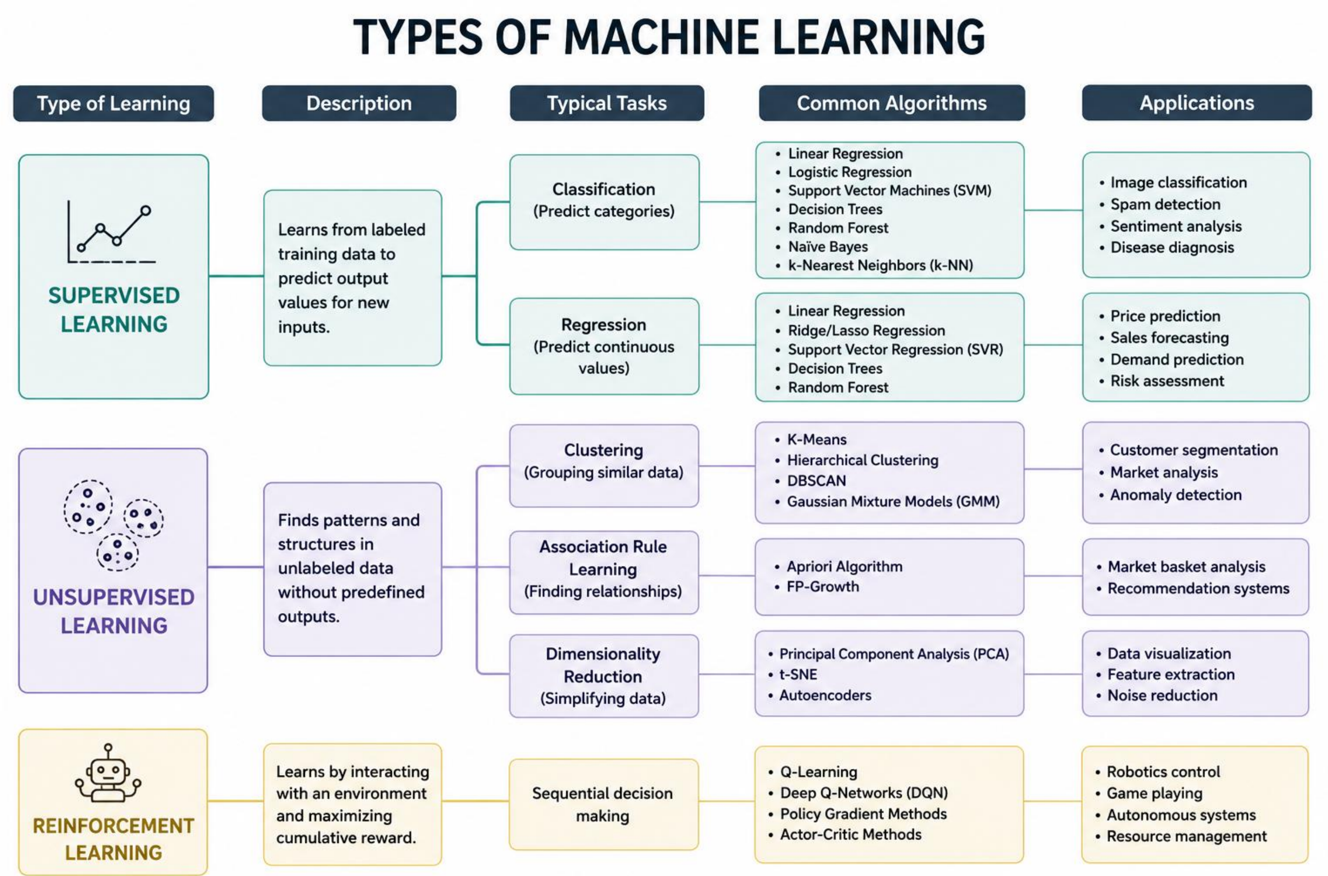

Figure 2. 2: Conceptual Overview of Machine Learning.

3.1 OVERVIEW OF MACHINE LEARNING TECHNIQUES

Artificial Intelligence, particularly through Machine Learning, employs various learning approaches to analyze data and identify significant patterns. Among these approaches, supervised learning and unsupervised learning represent the two main paradigms, differing in their training methods, data requirements, and practical applications

3.1.1 SUPERVISED LEARNING

Supervised Learning is considered one of the core paradigms of Machine Learning, where models are trained using labeled datasets in which each input feature is associated with a corresponding target output. The main goal of this approach is to establish a functional relationship between inputs and outputs, commonly represented as:

$$f(x)=y$$

This relationship enables the model to perform accurate predictions or classifications on previously unseen data. A supervised learning framework generally consists of four key components: input data (independent variables), output data (target variables), the learning algorithm itself such as Decision Trees, Linear Regression, or Support Vector Machines (SVM) and the training process, which iteratively adjusts model parameters using optimization techniques like Gradient Descent to minimize prediction errors. The supervised learning workflow is typically divided into two major stages. The first is the training phase, during which the model learns hidden patterns and relationships from the training dataset to build a predictive model. The second is the testing phase, where the trained model is evaluated using unseen data in order to assess its predictive capability and its ability to generalize effectively to real-world scenarios [14][15] [16].

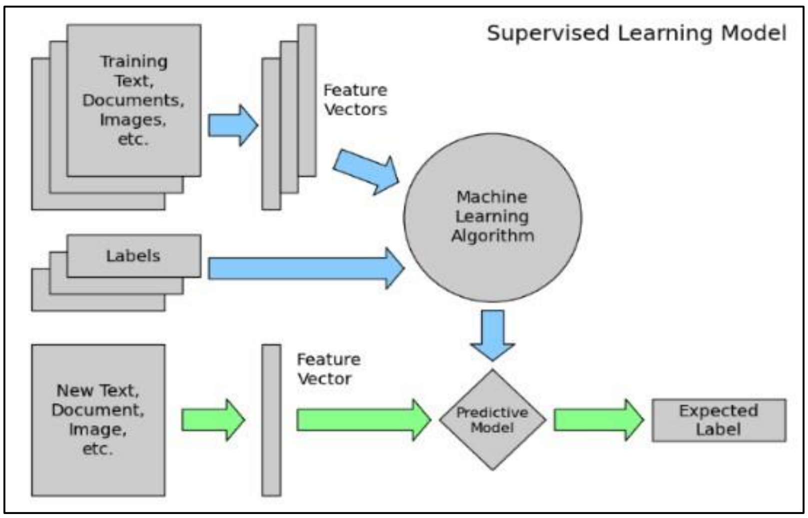

Figure 2. 3: Supervised Learning Model.

3.1.2 UNSUPERVISED LEARNING

Unsupervised learning is a core branch of machine learning that identifies hidden patterns and structures in unlabeled data without predefined outputs. Its process includes data preprocessing, applying unsupervised algorithms, and interpreting the results. Common applications include clustering and dimensionality reduction, which help simplify data while preserving important information. These methods enhance the understanding of complex datasets and support exploratory data analysis (EDA) and feature engineering before supervised learning [14].

3.1.3 ACADEMIC CONTEXT AND FRAMEWORK

According to Dramsch (2020) in "70 years of machine learning in geosciences," the integration of AI is not merely a trend but a fundamental shift in handling multi-modal geophysical data. Supervised learning serves as a "proxy-model" for computationally expensive physics-based simulations, such as using neural networks to approximate the Zoeppritz equations for AVO (Amplitude Variation with Offset) analysis. Conversely, unsupervised learning acts as a powerful exploratory tool for "big data" problems where human labeling is impossible, such as identifying patterns in massive passive seismic datasets or global climate models. The synergy between these methods allows for "Physics-Informed Machine Learning," which ensures that AI predictions remain consistent with the fundamental laws of thermodynamics and fluid dynamics [15].

3.2 MACHINE LEARNING ALGORITHM

3.2.1 LINEAR REGRESSION

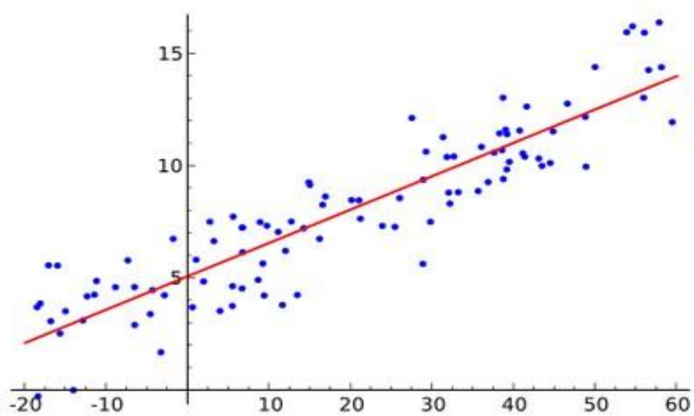

Figure 2. 4: Linear Regression Algorithm.

Linear Regression is a supervised machine learning technique widely employed for predictive modeling tasks. It aims to model the relationship between independent variables and a continuous dependent variable through a linear equation. During the training process, the algorithm estimates the optimal coefficients (weights) that minimize the difference between predicted and actual values in the dataset [16].

3.2.2 Decision Tree

A decision tree consists of nodes representing the output variable (y) used for prediction and classification tasks. By traversing the branches of the tree from the root to the terminal leaf nodes, the model determines the final category or predicted outcome. This approach is recognized for its simplicity, speed, and accuracy in solving a broad range of prediction problems [17].

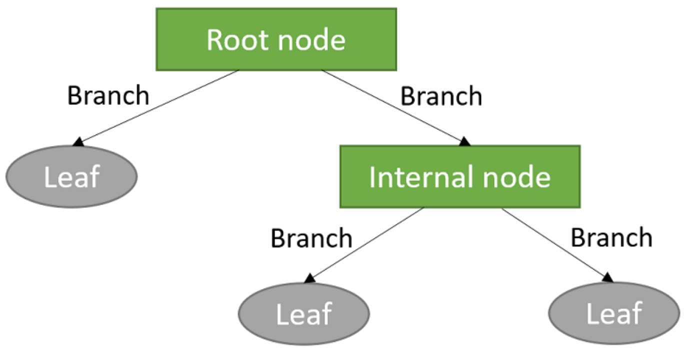

Figure 2. 5: Decision Tree Structure.

3.2.3 K-Nearest Neighbors (KNN)

The K-Nearest Neighbor (KNN) algorithm is an instance-based machine learning method used for prediction and classification tasks. It predicts the outcome of new data by comparing it with the most similar data points in the training dataset. In general, the accuracy of the model improves when a larger amount of training data is available [8].

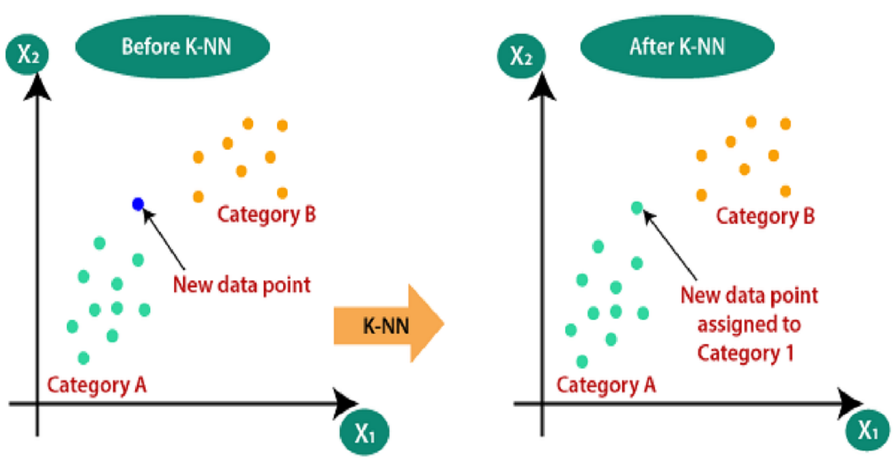

Figure 1. 6: Illustration of K-Nearest Neighbors Classification [9].

3.2.4 SUPPORT VECTOR MACHINE (SVM)

Support vector machines (SVMs) can be used to handle classification, regression, and outlier problems that are frequently encountered in supervised learning. The SVM is incredibly powerful on higher dimensional data, but has significant drawbacks when the data is very large and noisy [20].

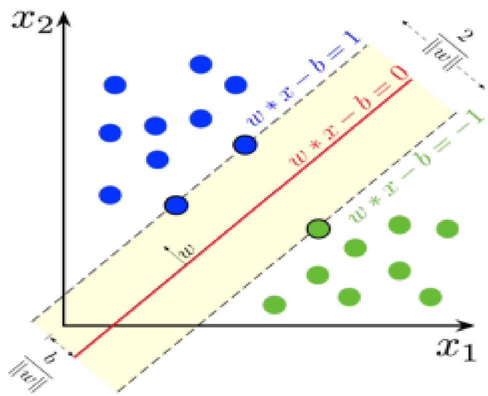

Figure 1. 7: Conceptual Illustration of SVM [20].

3.2.5 ENSEMBLE LEARNING ALGORITHMS

Ensemble learning algorithms are machine learning methods that improve accuracy by combining several models instead of using just one. The idea is that multiple models working together usually give better and more reliable results than a single model. The main types of ensemble methods are bagging, boosting, and stacking, which differ in how the models are built and combined. These techniques are commonly used to reduce overfitting and improve how well the model performs on new data [21] [22].

- **RANDOM FOREST**

Random Forest is an ensemble method that constructs multiple decision trees and aggregates their outputs. It handles high-dimensional, non-linear data well and reduces overfitting by combining diverse tree predictions [23].

- **EXTREME GRADIENT BOOSTING**

An optimized distributed gradient boosting library designed for high computational efficiency. It incorporates L1 and L2 regularization (Elastic Net) to mitigate overfitting and utilizes parallel processing to handle complex, large-scale datasets.

- **LIGHT GRADIENT BOOSTING**

A framework that employs histogram-based learning and a leaf-wise tree growth strategy. It is specifically engineered to manage high-dimensional data, significantly reducing memory consumption and training latency compared to traditional level-wise growth [24].

- **CATEGORICAL BOOSTING**

A specialized implementation designed to handle categorical features natively. It utilizes "Ordered Boosting" to overcome prediction shift and eliminates the need for extensive preprocessing techniques, such as One-Hot Encoding, while maintaining structural integrity in the data [24].

3.2.6 DEEP LEARNING

Neural Networks are a branch of machine learning designed to imitate how the human brain learns. They consist of interconnected nodes called neurons, where each neuron processes input data and passes the result to other neurons in the network. Artificial Neural Networks (ANNs) are the earliest form of neural networks and typically include one or more fully connected hidden layers between the input and output layers. These hidden layers learn patterns from the data using techniques such as back propagation [25].

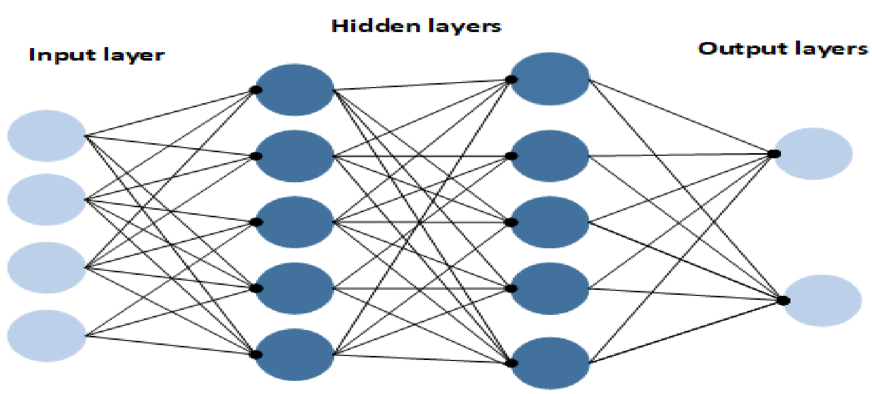

Figure 1. 8: Simple Architecture of an Artificial Neuron.

4 CONCLUSION

Machine learning methods are known for their ability to learn and generalize from data collected from different sources. This makes them suitable for handling environmental datasets that may be incomplete or have different levels of precision. Recent studies have shown that the integration of machine learning with remote sensing data improves prediction accuracy and helps in understanding complex environmental systems more effectively.

CHAPTER 3

METHODS AND MATERIALS

1 INTRODUCTION

Understanding the petrophysical and hydrodynamic characteristics of reservoir rocks is essential for petroleum engineering applications. These properties are commonly evaluated using well logging techniques and geophysical methods, which provide valuable insights into reservoir formations and fluid behavior [30].
Traditional methods for evaluating petrophysical properties in petroleum reservoirs commonly rely on laboratory measurements and empirical interpretations derived from well logging data. These approaches are widely used to characterize reservoir formations and estimate important subsurface parameters. However, such methods often present limitations in accurately representing in-situ reservoir conditions, since they may require transporting fluid or rock samples to the surface, which can alter their original properties. In addition, these procedures are generally time-consuming, expensive, and operationally demanding. This chapter focuses on some of the most widely used machine learning approaches, including margin-based and instance-based algorithms such as Support Vector Machines (SVM) and K-Nearest Neighbors (KNN), as well as ensemble learning methods like Random Forest (RF), CatBoost, and Extreme Gradient Boosting (XGB). These data-driven techniques have demonstrated strong performance in modeling complex subsurface relationships [31] [32].

In particular, they are capable of delivering continuous petrophysical properties along the wellbore, thereby reducing or eliminating the need for extensive laboratory analyses and repeated sampling operations.

2 OVERVIEW OF WELL LOGGING

2.1 DEFINITION

Similar to traditional methods and direct measurements, there are also indirect methods that help in evaluating and characterizing petroleum reservoirs more effectively and accurately. Well logging, or borehole logging, refers to the systematic recording of geological formations encountered during drilling. This record, known as a “well log,” is generated using downhole instruments (sondes) that are lowered into the wellbore. These tools measure and record variations in the physical properties of the surrounding rock formations and the fluids they contain, either continuously or at specific depth intervals [33] [34].

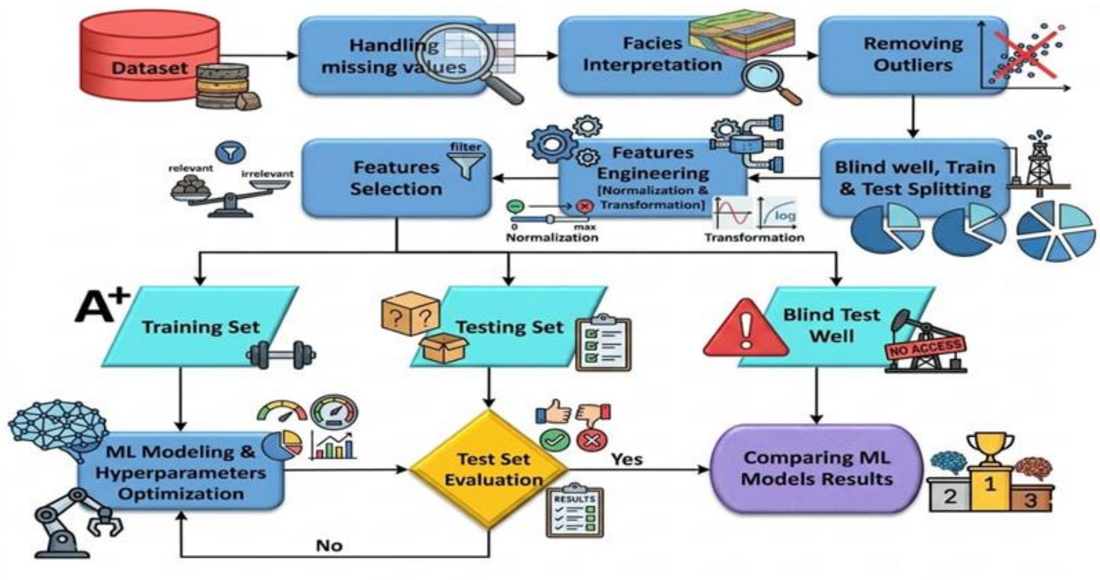

Figure 3. 1: General Workflow for Training Machine Learning Models.

2.2 COMMON WELL LOGS

A broad range of well logging tools has been developed, with each tool designed to measure specific physical properties of the subsurface formations. The selection of logging measurements in a well depends on the objectives of the logging program, the expected lithology, and economic factors. Among the most widely used and essential logs are the following [34]:

- **Gamma Ray (GR) Log:** This log measures natural gamma radiation from rock formations. Shale's usually give high readings because they contain radioactive elements like potassium, thorium, and uranium. Clean formations such as sandstone and carbonate give low readings. It is mainly used to identify lithology (shale vs. non-shale) and for stratigraphic correlation.
- **Resistivity logs**: Measure how much a formation resists the flow of electrical current. Hydrocarbon zones usually have high resistivity, while water-bearing formations, especially with salty water, have low resistivity because they conduct electricity. These logs are used to identify hydrocarbon zones and estimate water saturation. Different tools like induction and laterologs are used depending on borehole conditions.
- **Density Log (RHOB):** This log works by sending gamma rays into the formation and measuring the scattered signals that return. The response depends on the electron density of the rocks, which is closely related to bulk density. It is mainly used to estimate porosity and help identify lithology.

- **Neutron Porosity Log (NPHI):** This tool emits high-energy neutrons into the formation and measures the returned neutron or gamma radiation. It mainly responds to hydrogen content, which is mostly found in formation fluids such as water or hydrocarbons. Therefore, it is widely used to estimate formation porosity.
- **Sonic Log (DT):** This log measures the time taken by sound waves to travel through a formation. The travel time depends mainly on rock type and porosity. It is used to estimate porosity, identify lithology, evaluate rock mechanical properties, and help in seismic correlation.
- **Caliper Log:** This log measures the size (diameter) of the borehole. Changes in hole size can indicate issues such as shale collapse, formation washouts, or mud cake buildup in permeable zones. It is also used to check borehole conditions and the quality of other logging measurements [33].

Table 3. 1: Overview of Well Log Measurements, Units, and Geological Significance.

Feature	Description	Units	Usage
DEPTH	Measurement depth	m / ft	Reference for all logs
GR	Gamma ray intensity	gAPI	Indicates shale or radioactive content
DT	Acoustic slowness	$\mu s/ft$	Helps estimate porosity and lithology
CALI / CALX	Borehole diameter	inches	Detects borehole washout
NPHI / CNC	Neutron porosity	m^3/m^3	Sensitive to hydrogen used for porosity
RHOB	Bulk density	g/cm^3	Formation density; lithology analysis
PE	Photoelectric factor	barns/e	Indicates mineral composition
RT / RT10–90	Formation resistivity	ohm·m	Distinguishes fluids and lithology
RXO	Micro-resistivity	ohm·m	Reflects flushed zone near borehole
K / POTA	Potassium	%	Shale and clay indicator
TH / THOR	Thorium	ppm	Radioactive marker; shale content
U / URAN	Uranium	ppm	Indicates organic material or shale
M2R1–M2R9	Multi-depth resistivity	ohm·m	Resistivity at several radial depths
TEMP	Formation temperature	°C	Affects log tools and calibration
SALINITY	Water salinity	ppm	Impacts resistivity and saturation analysis

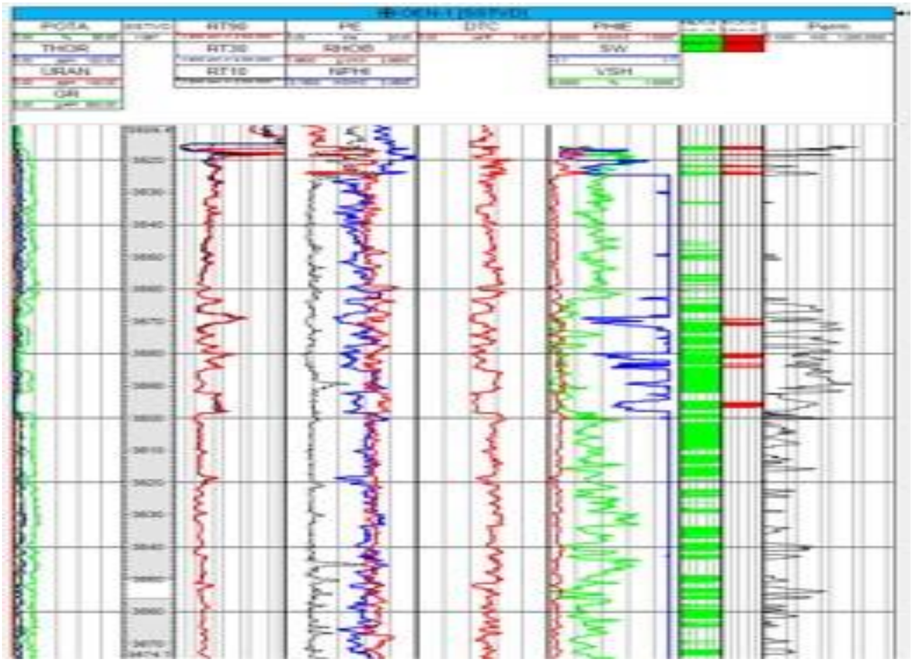

Figure 3. 2: Example of well logs.

2.3 IMPORTANCE OF WELL LOGGING

- **Formation evaluation**: uses well logs continuous measurements taken along a borehole to calculate essential rock and fluid properties [33].
- **Characterization of Reservoirs:** Geoscientists can construct detailed three-dimensional models of reservoirs by combining log data from several wells. These models help them understand the reservoir's shape, internal variations, and how fluids are distributed throughout it [35].
- **Detection and Hydrocarbons Estimation:** Resistivity logs and porosity logs are essential tools for locating zones that contain oil or gas, and for estimating the total volume of hydrocarbons stored underground [36].
- **Geological Correlation:** Gamma Ray (GR) and Spontaneous Potential (SP) logs are widely used to match rock layers between different wells. This helps geoscientists create maps that show the structure and sequence of underground formations [37].
- **Geomechanical Analysis:** Sonic logs and density logs are used to calculate the mechanical properties of rocks. This information is essential for ensuring wellbore stability, designing hydraulic fracturing operations, and studying ground subsidence [38].
- **Seismic Interpretation:** Well log data particularly density and sonic logs is used to create synthetic seismograms. These synthetic traces help calibrate and interpret surface seismic data, connecting detailed well-scale measurements with broader seismic-scale exploration [39].

2.4 PETROPHYSICAL PARAMETERS

Geophysics uses physical methods to study the Earth's hidden interior layers. In the oil and gas industry, these methods help locate potential reservoirs, but the presence of hydrocarbons can only be confirmed by drilling a well. Once a well is drilled, well logging is used to measure the specific rock and fluid properties of the reservoir to determine if it is productive [40].

- **Porosity**: is the measure of the empty space (voids) within a rock compared to its total volume. It is expressed as either a decimal or a percentage and is typically symbolized by the Greek letter phi (ϕ) [40].
- **Fluid Saturation (S_w, S_h):** This represents the portion of a rock's total pore space filled by a particular fluid. For example, water saturation (Sw) measures the fraction of pores containing water, while hydrocarbon saturation (Sh) measures the fraction containing oil or gas [41].
- **Formation Water Resistivity (Rw):** This is the measure of how much the water trapped inside the rock pores resists the flow of electricity. It is a critical value used in calculations to determine how much water is present in a reservoir based on electrical log data [42].
- **Shale Volume (Vsh):** This represents the portion of the rock made up of shale or clay minerals. Calculating this is crucial because a high concentration of clay can block pore spaces and reduce the overall quality and flow capacity of a reservoir [43].
- **Lithology:** This describes the physical makeup of the rock, including its mineral composition and texture (such as sandstone, limestone, or shale) [44].
- **Permeability (k):** A measure of the ability of a rock to transmit fluids through its interconnected pore network. It is critical for producibility but is generally not measured directly by logs, rather inferred or estimated [45].

3 THE CONTRIBUTION OF SONIC LOG IN PETROLEUM GEOSCIENCE

The sonic log measures the transit time interval (Δt), defined as the time required for a compressional acoustic wave to travel through one foot of formation. This interval is the reciprocal of compressional wave velocity and varies according to the formation's lithology and porosity [46].

In petrophysics, sonic logs assist in porosity estimation using the Wyllie time-average equation and help evaluate rock mechanical properties such as Young's modulus and Poisson's ratio2. In petroleum engineering, compressional sonic logs aid in wellbore stability analysis, fracture identification, and real-time drilling optimization.

$$\Delta tlog = \phi \Delta tf + (1 - \phi)\ \Delta tma \tag{3.1}$$

Rearranging for porosity:

$$\phi S = (\Delta tlog - \Delta tma) / (\Delta tf - \Delta tma) \tag{3.2}$$

Where Δtma is the matrix transit time (e.g., 55.5 µs/ft for sandstone, 47.5 µs/ft for limestone) and Δtf is the fluid transit time (e.g., 189 µs/ft for water). This equation performs best in consolidated formations with intergranular porosity; for vuggy or fractured porosity, alternative models such as Raymer-Hunt-Gardner are more suitable [46].

The sonic logging tool functions by measuring the travel time of an elastic wave pulse emitted from a transmitter to a receiver, both mounted on the tool body. The transmitted pulse is characterized by high amplitude and a short duration. As this energy propagates through the formation, it undergoes both dispersion the spreading of energy in time and space and attenuation, which is the loss of energy as it is absorbed by the rock (Schlumberger, 1989).

Upon arrival at the receiver, the signal is complex because different wave types travel at varying velocities or follow distinct pathways through the rock. To prevent the high-amplitude transmitted pulse from interfering with the measurement and to protect the sensitive receiver, the system is electronically masked at the start (t = 0).

The recorded wave train typically arrives in the following sequence:

- **Compressional Wave (P-wave):** This is the longitudinal pressure wave. It is the fastest to arrive, typically appearing with a relatively low amplitude.
- **Shear Wave (S-wave):** This transverse wave generally arrives after the P-wave. It travels at a slower velocity but is typically recorded with a higher amplitude.

4 ADDRESSING AND SYNTHESIZING MISSING DATA IN WELL LOGS

Well log datasets are frequently incomplete, containing missing values or entire missing log curves for certain depth intervals or wells. The missing data can occur for several causes, like tool failure, formation conditions, and human factors [47]. This missing data can significantly hinder petrophysical analysis, reservoir characterization, and the application of data-driven techniques [48][49].

4.1 CAUSES OF MISSING DATA

Well log datasets are seldom complete. Gaps may appear as isolated missing values or as entire missing curves across certain depth intervals or wells. These omissions can seriously impede petrophysical interpretation, reservoir characterization, and the application of data driven techniques [50].

- **Operational issues:** Logging runs may be shortened or abandoned when borehole conditions become unfavourable such as tight spots or bridging or when time and cost limits are reached [52].

- **Borehole conditions:** Severe washouts and rugosity can make some tool responses unreliable, so those readings are flagged and effectively removed from the dataset [54].
- **Tool malfunction or failure:** Downhole instruments can fail because of high temperature, pressure, or mechanical wear, producing gaps in the acquired data [51].
- **Selective logging programs:** Operators often run advanced logs only in key wells or over zones of primary interest, leaving other intervals without measurements for budgetary or strategic reasons [53].
- **Depth mismatch and splicing:** When several logging runs are merged, small depth discrepancies between passes can create artificial gaps if the depths are not perfectly reconciled [55].
- **Data processing and transmission errors:** Mistakes made while recording, transmitting, or processing the data can corrupt or erase segments of the log [49].

5 THE MODEL TRAINING PIPELINES : KEY STAGES AND PRACTICES

5.1 DATA COLLECTION

Collecting good, representative data is essential. Quality data enhances model accuracy, generalization, and helps minimize biases [13].

5.2 DATA PREPROCESSING

Data preprocessing encompasses managing missing values, normalization, encoding, and converting raw data into a model ready format. This foundational step is crucial for ensuring model dependability [13].

5.2.1 DATA TRANSFORMATION

Data transformation converts raw information into formats optimized for machine learning, encompassing operations such as scaling, normalization, and feature engineering to enhance model compatibility and performance [56].

5.2.2 SPLITTING DATA

Partitioning data into training, validation, and test subsets enables unbiased evaluation of model performance and provides reliable estimates of generalization capability on unseen data [21].

5.3 MODEL SELECTION

Selecting an optimal model requires comparing algorithms and configurations through performance metrics such as accuracy, precision, or F1-score, typically employing cross-validation to assess generalization. This process aims to strike a balance between predictive power and interpretability based on specific application needs [13].

5.4 MODEL TRAINING AND VALIDATION

The model extracts patterns from training data, while validation serves to refine hyperparameters and mitigate overfitting during the learning process [25].

5.5 MODEL EVALUATION

The imputation and prediction techniques performance quality were assessed using four standard regression error metrics, computed only over the artificially masked values:

- **Root Mean Squared Error (RMSE):** The square root of the calculated average squared deviation between actual and imputed/predicted value, offering an interpretable error scale.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (3.3)$$

- **Mean Absolute Error (MAE):** Captures the average absolute difference between predicted and observed values, offering a robust assessment less sensitive to outliers.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3.4}$$

- **Coefficient of Determination (R^2):** Evaluates the proportion of variance in the true values accounted for by the imputed values, indicating the overall goodness of fit.

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \tag{3.5}$$

These metrics provide complementary insights into the accuracy, robustness, and consistency of each imputation method. R^2 values near 1 indicate strong agreement with the ground truth, while lower RMSE and MAE indicate more precise reconstructions.

- **Spearman Correlation Coefficient:** works by converting the actual physical values of your logs into ranks (1st, 2nd, 3rd, etc., from lowest to highest value) and then calculating the correlation based on those ranks [68].

6 SOFTWARE USED

6. 1 GEOXMATE

GeoXMate is a comprehensive Python-based graphical user interface (GUI) software developed to simplify and accelerate the use of machine learning techniques in petroleum geoscience and subsurface analysis. The software was created to address one of the major challenges faced by geoscientists although machine learning has become increasingly important for reservoir characterization, lithofacies classification, seismic interpretation, and petrophysical prediction, many professionals in geology and geophysics lack the programming skills required to implement advanced machine learning workflows. Including LAS well logs, SEGY seismic data, well deviation surveys GeoXMate overcomes this limitation by providing a fully integrated and user-friendly environment where users can perform all stages of the machine learning workflow without writing code. The platform combines data loading, preprocessing, visualization, feature engineering, model training, evaluation, and deployment into a single coherent system, which improves workflow efficiency, transparency, and reproducibility [57].

6. 2EXCEL

Microsoft Excel is a widely used spreadsheet application developed by Microsoft, designed for organizing, analyzing, visualizing, and calculating data within a grid-based interface of rows and columns. It offers an extensive library of built-in functions for diverse computational tasks.

7 CONCLUSION

The chapter details the paradigm shift in petroleum geoscience from traditional, destructive, and costly laboratory reservoir testing to advanced, data-driven machine learning (ML) frameworks. By utilizing indirect measurements from subsurface well logging tools, these modern techniques fill data gaps, optimize workflows, and generate continuous, highly accurate profiles of critical reservoir parameters.

CHAPTER 4

RESULTS AND DISSCUTION

1. INTRODUCTION

The "Oglet EN Nasser" field is located in the Amguid-Messaoud basin in Algeria, north of Hassi Messaoud field. Although Sonatrach has drilled many wells there and has high quality seismic data, there is often a difference between the depths predicted by seismic maps and the actual depths found during drilling. To solve this, geophysicists use "sonic logs" to create synthetic seismograms. This helps them improve their velocity models and better understand the underground layers. Additionally, because some formations in this field have low productivity, hydraulic fracturing is often necessary. Accurate compressional wave velocity (Vp) data is essential for modeling, predicting shear waves, and planning these fracturing operations.

Usually, compressional wave velocity is measured directly using wireline tools or estimated using standard formulas like the Wyllie or Raymer-Hunt-Gardner models. However, direct measurements are very expensive and require extra time on the drilling rig. Keeping the well open for these measurements also increases the risk of the wellbore collapsing or damaging the rock formations. To address this, we can use machine learning to predict missing sonic logs by finding patterns in other petrophysical data. This is a faster and more cost effective way to estimate velocities where direct data is missing.

This chapter presents our results in predicting compressional travel time (DTC) using four machine learning models: Random Forest, XGBoost, CatBoost, and Artificial Neural Networks. We trained these models using cleaned data from five different wells. We then tested the models by predicting the DTC curve for a sixth, "blind" well. Our results show that the Random Forest algorithm was the most accurate and performed better than the other models.

2. METHODOLOGY

The methodology involves collecting available well log data from five wells within the study area, covering the interval from the Lower Triassic Clayey Sandstone Formation (TAGI) formation to the top of the El Gassi Clay (AEG) formation. Following data collection, data cleaning and preprocessing are conducted, which includes handling missing values and resolving any data inconsistencies.The next step is feature selection, where the most critical variables are identified, such as: Gamma Ray (GR), Resistivity, Density (RHOB), Neutron, Photoelectric Factor (PEF), and Compressional Slowness.

Additionally, feature engineering techniques such as data transformation and normalization are applied where necessary. The selected features are then split, utilizing data from the five wells : ['OEN-a' 'OEN-b' 'OEN-c' 'OEN-d' 'OEN-e'] for the training and testing sets, while a sixth well is set aside as an independent blind test well.

Several machine learning regression algorithms are explored and tested, including Random Forest, CatBoost, XGBoost, and Artificial Neural Networks (ANN). Each model is trained using the same training dataset, with hyperparameter tuning performed to achieve the highest possible predictive performance. Finally, the models' performance is evaluated using appropriate validation metrics on both the test set and the independent well.

2.1 DATA COLLECTION AND PREPROCESSING

➢ HANDLING MISSING VALUES

Machine learning models generally require complete datasets, so missing values in both input features and target variables need to be handled before training. In this study, missing data were filled using the average of neighboring values instead of removing incomplete records, in order to preserve as much information as possible. This method is reasonable because the data were sampled at a high resolution of 0.5 ft, where geological properties are not expected to change significantly over such small intervals. In addition, missing values occurred in very short sequences (no more than two consecutive points), which makes interpolation effects minimal [21].

The proportion of missing data was also very low (about 10 % to 12 %), so the risk of introducing bias through imputation is limited. Overall, this simple interpolation approach maintains data continuity, preserves geological trends, and supports reliable model training [58].

Table 4.1: Initial vs. Cleaned Dataset Breakdown and Feature Retention Percentages.

Feature	Data Type	Initial Dataset (N=21,294)	Cleaned (N=18,623)	Missing Values (NaN)	Feature Retention (%)
Depth	float64	21,294	18,623	0	100.0%
MD	float64	21,294	18,623	0	100.0%
DTC	float64	20,621	18,623	673	96.8%
GR	float64	20,909	18,623	385	98.2%
NPHI	float64	19,203	18,623	2,091	90.2%
PE	float64	19,300	18,623	1,994	90.6%
RHOB	float64	19,300	18,623	1,994	90.6%
RT90	float64	20,154	18,623	1,140	94.6%
Count	—	—	**18,623**	**2,671 (Rows Dropped)**	**87.45%**

➢ DESCRIPTIVE STATISTICS

- **Univariate Statistical Analysis**

Prior to analysis, each well log parameter was assessed for data quality, including the presence of missing values and outliers. The dataset was generally of good quality and suitable for analysis, requiring only minor preprocessing through the removal of a few outlier values [59] [60].

The dataset from Table 4.2 contains severe, non-physical outliers such as impossible bulk density (1.76 to 4.63 g/cm^3), an unphysically high neutron porosity (0.68), and extreme spikes in Gamma Ray (1649.37 API) and Resistivity (9380.61 ohm.m) that are driven by borehole washouts, tool malfunctions, or operational noise rather than true geology. Because these anomalies violate domain constraints, they must be filtered or imputed using a Caliper log prior to training to prevent biased machine learning parameters and ensure reliable model predictions [62].

Table 4.2: Descriptive statistics of the original input features and target variable

Feature	Count	Mean	Std. D	Min	Q1	Q2	Q3	Max
Depth	21,294	-3,668.19	139.71	-4,024.80	-3,767.50	-3,661.10	-3,554.60	-3,385.20
MD	21,294	3,807.23	139.77	3,526.20	3,693.40	3,799.90	3,906.30	4,162.80
DTC	20,621	72.35	11.29	42.77	63.47	70.47	80.01	138.80
GR	20,909	94.48	125.69	6.68	30.37	77.48	112.44	1,649.37
NPHI	19,203	0.21	0.10	-0.03	0.14	0.22	0.28	0.68
PE	19,300	4.86	2.42	1.87	3.50	4.11	5.03	15.41
RHOB	19,300	2.62	0.31	1.76	2.56	2.62	2.67	4.63

A proactive step, we observed from Figure 4.1-a- the presence of outlier values. Therefore, we applied a multivariate analysis between the parameters. Given the importance of the compressional velocity log, we selected it as an example to illustrate its relationship with another parameter (NPHI), as well as the method used to remove the outliers and their impact on the models.

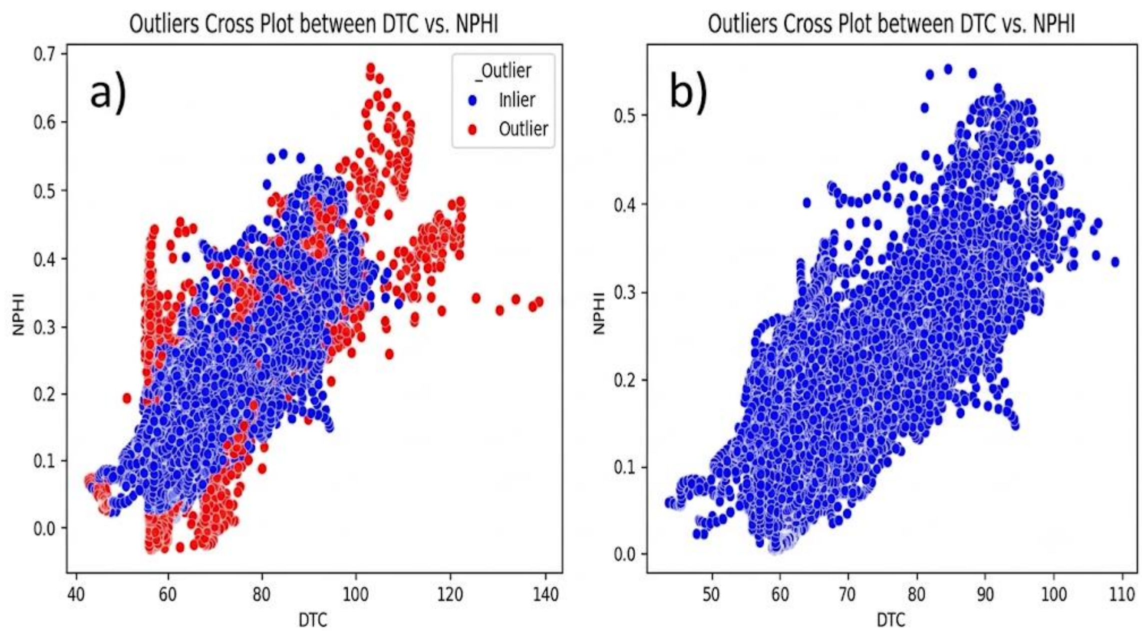

Figure 4.1: Cross-plot of Compressional slowness (DTC) versus Neutron Porosity (NPHI).

To improve data quality, the Isolation Forest algorithm was applied to detect and remove outliers from the well-log dataset. This tree-based anomaly detection method isolates abnormal observations through random recursive partitioning, making it efficient for large and high-dimensional datasets [61].

Table 4.3: Descriptive statistics of input features and target variable.

Feature	Count	Mean	Std. D	Min	Q1	Q2	Q3	Max
Depth	16,760.00	-3,690.64	118.51	-4,010.10	-3,765.20	-3,679.70	-3,595.50	-3,486.60
MD	16,760.00	3,829.52	118.69	3,627.60	3,734.30	3,818.50	3,904.00	4,148.10
DTC	16,760.00	72.08	10.05	43.59	63.73	70.33	79.83	108.91
GR	16,760.00	84.63	58.31	9.00	36.40	85.70	112.34	609.09
NPHI	16,760.00	0.21	0.09	0.01	0.14	0.22	0.27	0.55
PE	16,760.00	4.38	1.72	1.87	3.42	4.01	4.74	13.47
RHOB	16,760.00	2.62	0.08	2.20	2.57	2.62	2.67	3.76
RT90	16,760.00	15.10	48.70	0.58	2.15	3.89	8.04	1,344.72

The median neutron porosity value is 0.22 (22%), which represents good reservoir porosity. Following data preprocessing and outlier removal, the porosity distribution became more consistent and geologically realistic compared to the original dataset, enhancing the reliability of reservoir quality assessment and fluid storage potential evaluation [62].

- **Bivariate Statistical Analysis**

The figure 4.2 is shown, Pearson correlation analysis conducted on the combined dataset from all wells indicates that the compressional sonic transit time (DTC) is strongly positively correlated with neutron porosity (NPHI, $r_p = 0.7$) and gamma ray (GR, $r_p = 0.6$). This behavior is consistent with the slower propagation of compressional waves in porous and clay-rich formations. In contrast, DTC exhibits a moderate negative correlation with bulk density (RHOB, $r_p = -0.3$), reflecting the tendency of denser rocks to transmit acoustic waves more rapidly [63].

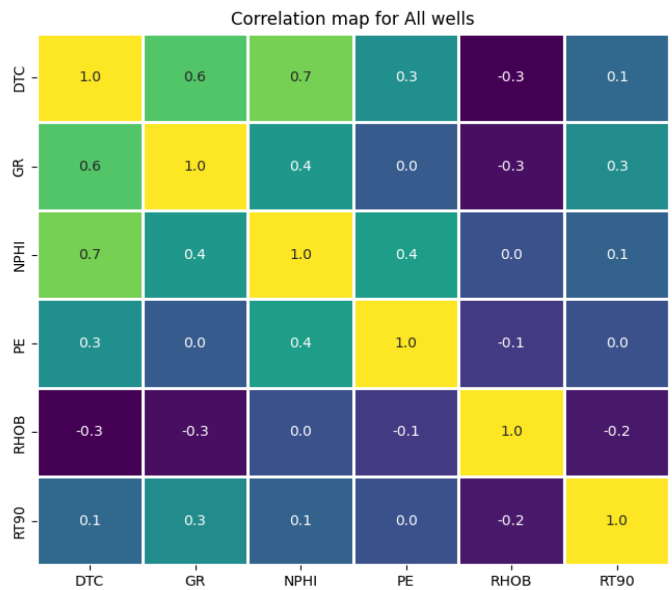

Figure 4. 2:Pearson correlation matrix.

The figure 4.3 is shown, Spearman correlation analysis indicates that compressional sonic transit time (DTC) has strong monotonic relationships with several well-log variables. DTC shows strong positive correlations with neutron porosity (NPHI) and gamma ray (GR) ($r_s = 0.7$), while deep resistivity exhibits a strong negative correlation ($r_s = -0.5$). These relationships reflect the influence of pore fluids, clay content, and reservoir hydrocarbon saturation on acoustic wave propagation [64].

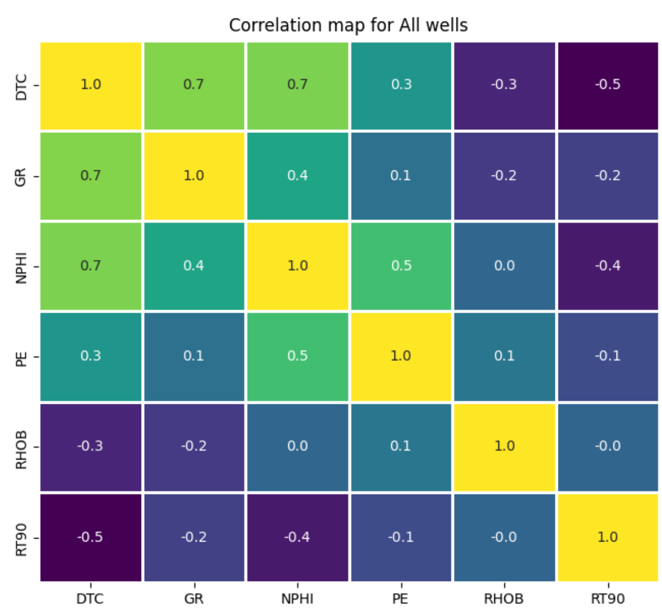

Figure 4. 3:Spearman correlation matrix.

Pearson correlation measures the strength of a linear relationship between two continuous variables, assuming normal distribution and sensitivity to outliers. In contrast, Spearman correlation assesses monotonic relationships based on ranked data, making it suitable for non-normally distributed datasets, ordinal variables, and data containing outliers, including non-linear relationships commonly observed in well-log interpretations such as resistivity–porosity interactions [65].

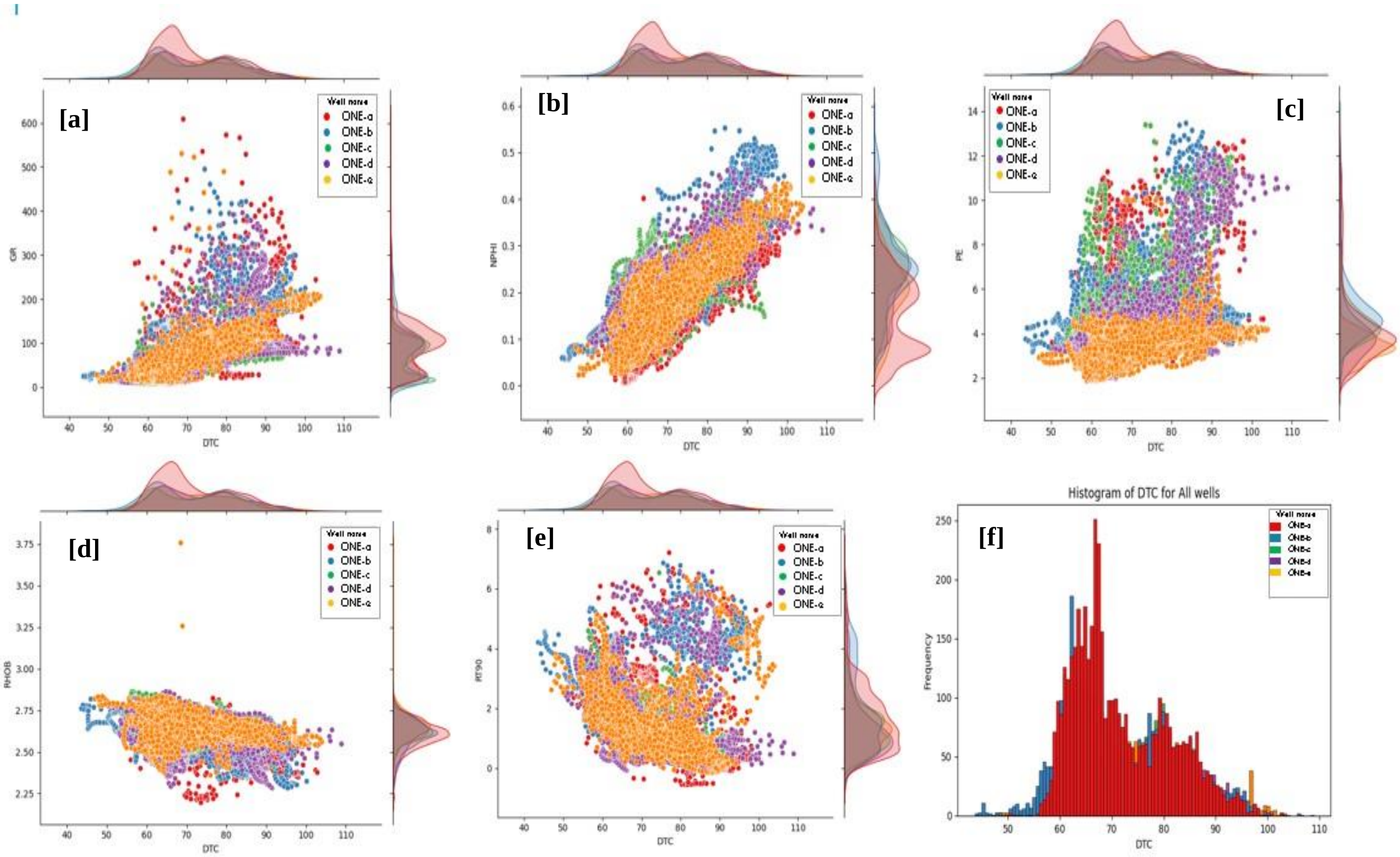

Figure 4. 4: Cross plot of measured DTC versus (a) GR, (b) NPHI, (c) PE, (d) RHOB, (e) Res 90, (f) DTC (Histogram)

In figure 4.4, the exploratory plots indicate a strong positive linear relationship between neutron porosity (NPHI) and compressional sonic transit time (DTC), mainly driven by their shared dependence on porosity. In contrast, deep resistivity (RT90) and bulk density (RHOB) show more complex, non-linear behaviors influenced by outliers and lithological variations. This explains why Pearson correlation may fail to capture the true inverse relationship of resistivity with DTC, which is better revealed by Spearman correlation (r_s = −0.5). The multimodal distribution of DTC further highlights the need for proper data cleaning and outlier handling prior to machine learning modeling to ensure robust and accurate predictions.

2.2 MACHINE LEARNING MODELS TRAINING AND VALIDATION

2.2.1 DATASET SPLITTING

Four machine learning models were used to predict compressional sonic transit time (DTC), including ensemble methods (Random Forest, XGBoost, CatBoost) and artificial neural networks (ANN). Random Forest uses bagging, where multiple decision trees are trained independently on random subsets of data and averaged to improve stability and reduce overfitting. In contrast, XGBoost and CatBoost are boosting-based methods that build trees sequentially to correct previous errors, offering high accuracy and efficiency, especially for large and complex datasets. Finally, ANN are capable of capturing complex nonlinear relationships through multiple hidden layers, making them effective for modeling highly complex geophysical data [23].

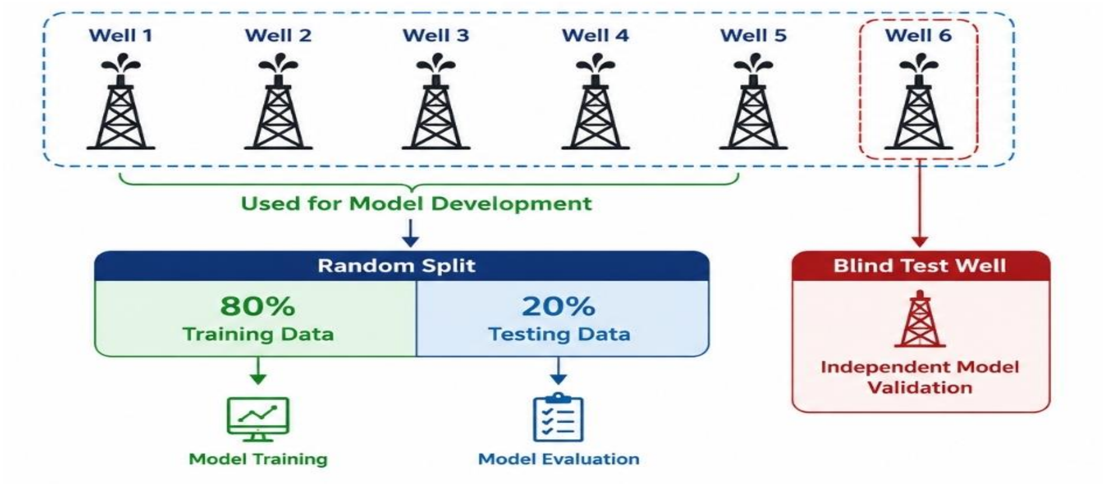

Figure 4. 5:Schematic illustration of the dataset partitioning strategy.

2.2.2 HYPERPARAMETER TUNING

Hyperparameter tuning is a key step in machine learning model development because it improves accuracy, stability, and generalization on unseen data. Although it can be computationally expensive, it is essential for achieving optimal model performance. In this study, Grid Search was used to optimize the hyperparameters of all models by systematically testing predefined parameter combinations and selecting the best performing configuration based on evaluation metrics. The optimal settings (Table 4.4) were then applied for final training and evaluation, resulting in improved predictive performance and better generalization to unseen wells [66] [67].

Table 4. 3: Optimal values of key hyperparameters across all algorithms.

Algorithm	Optimized Hyperparameters		
Random Forest	n_estimators = 100	min_samples_split = 2	min_samples_leaf = 1
XGBoost	n_estimators = 1000	learning_rate = 0.1	max_depth = 10
CatBoost	n_estimators = 100	learning_rate = 0.1	max_depth = 10
ANN	5 hidden layers	optimizer = Adam	learning_rate = 0.1

2.2.3 EVALUATION METRICS

Table 4. 5 presents a comparative analysis of the statistical performance metrics among the different machine learning algorithms across the training, testing, and blind test well datasets, aimed at identifying the model with the highest accuracy and generalization capability. Furthermore, Figures 10 and 11 display the scatter cross plots comparing the measured compressional sonic transit time (DTC) log values against the predicted values generated by the developed models for the training and testing datasets, respectively."

Table 4. 4: Performance evaluation metrics for training, testing, and blind well sets across all algorithms.

Algorithm	Set	MSE	RMSE	R^2
Random Forest	Training	0.874	0.935	0.991
	Testing	5.790	2.406	0.942
	Blind well	18.022	3.645	0.836
XGBoost	Training	1.281	1.131	0.987
	Testing	6.183	2.486	0.939
	Blind well	17.591	4.194	0.840
CatBoost	Training	7.128	2.670	0.929
	Testing	8.112	2.848	0.920
	Blind well	12.804	3.578	0.884
ANN	Training	15.007	3.874	0.851
	Testing	15.172	3.895	0.850
	Blind well	21.475	4.634	0.805

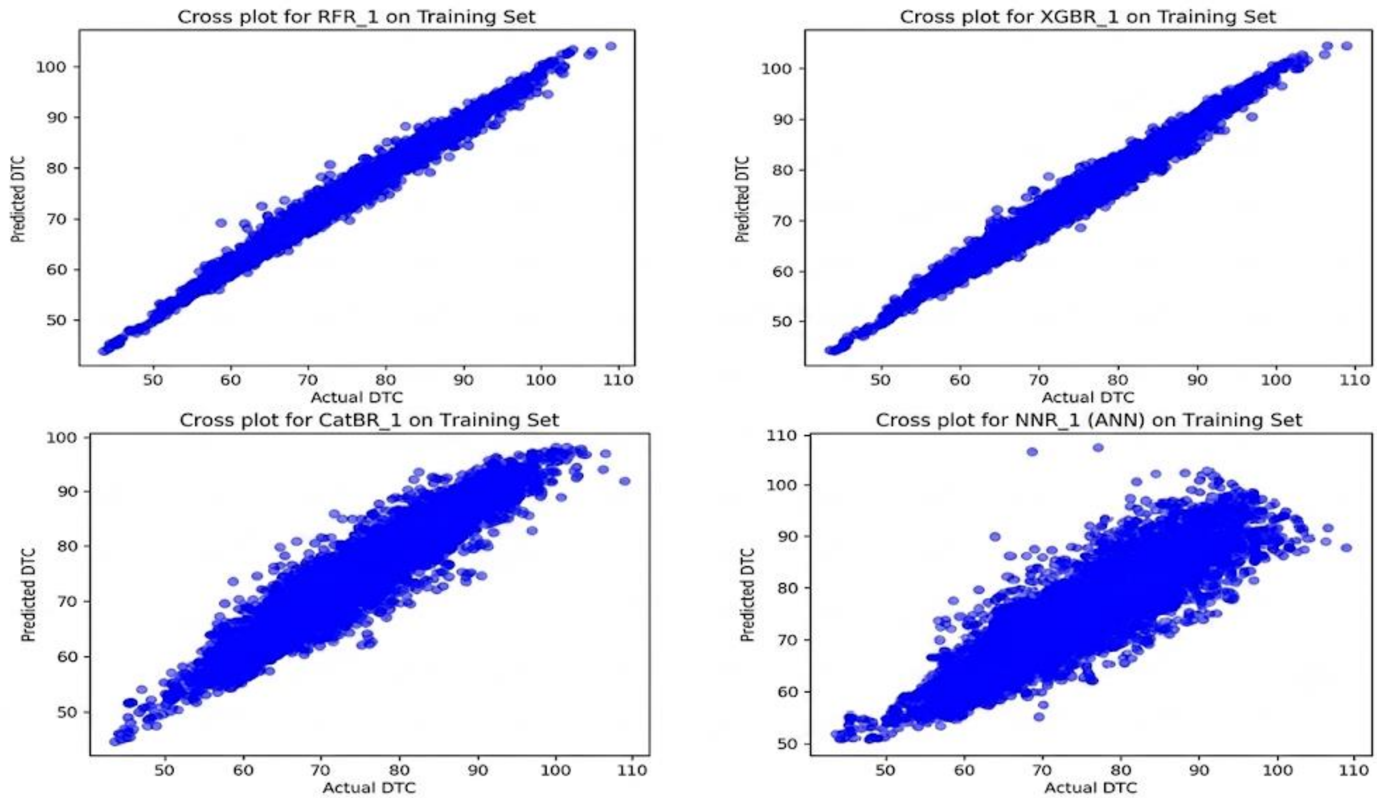

Figure 4. 6: Train set cross plots comparing original logs with synthetic logs generated by other models

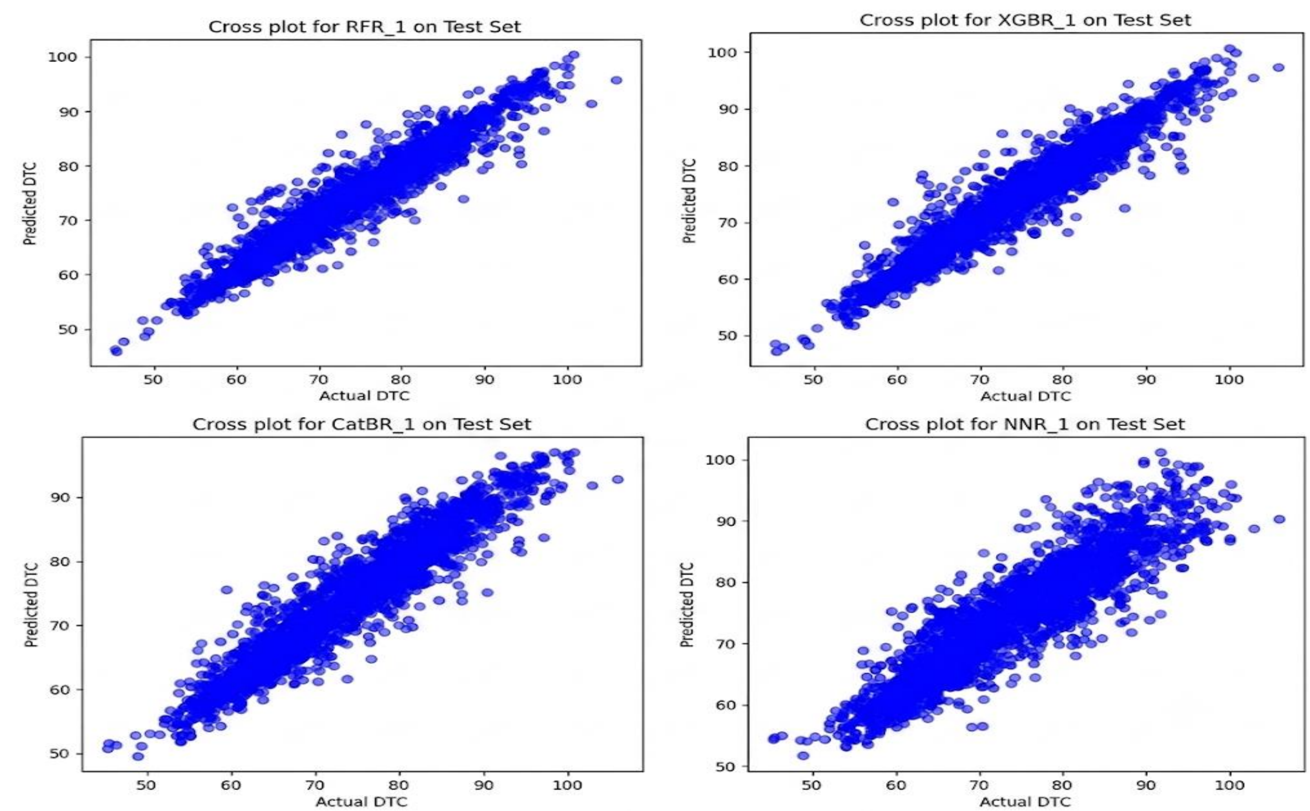

Figure 4. 7: Test set cross plots comparing original logs with synthetic logs generated by other models

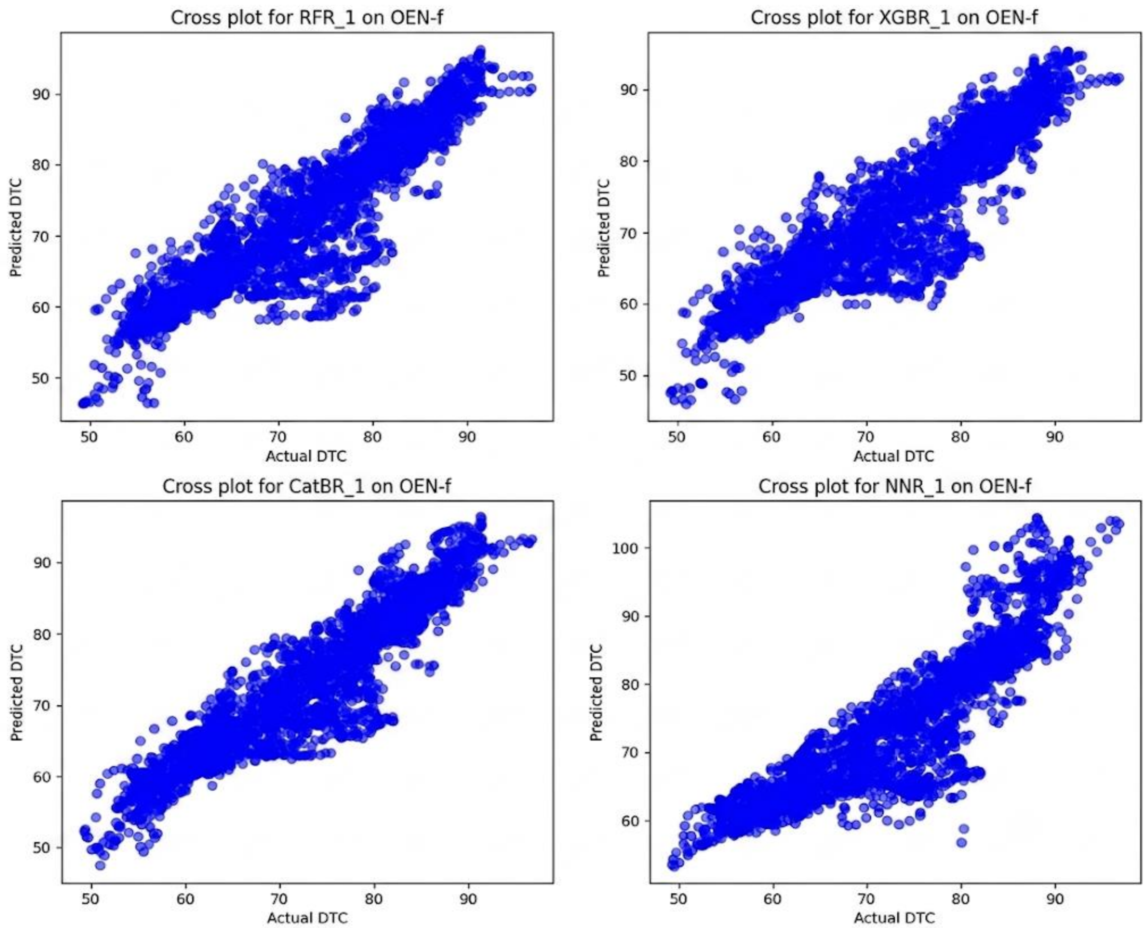

Figure 4. 6: Blind well plots comparing original DTC sonic logs with synthetic logs generated by other models.

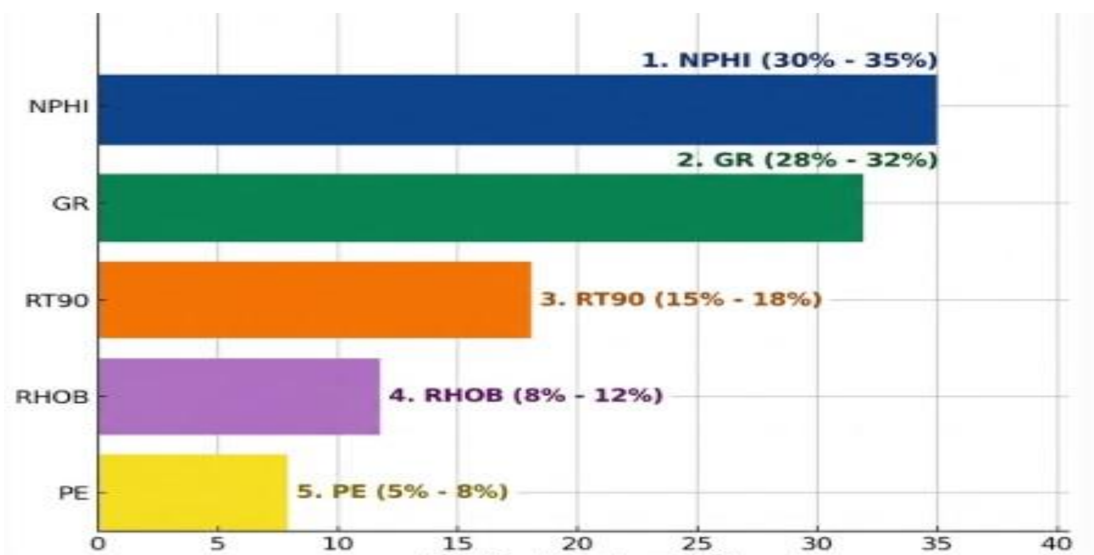

Figure 4. 7: Feature importance analysis for the RF model.

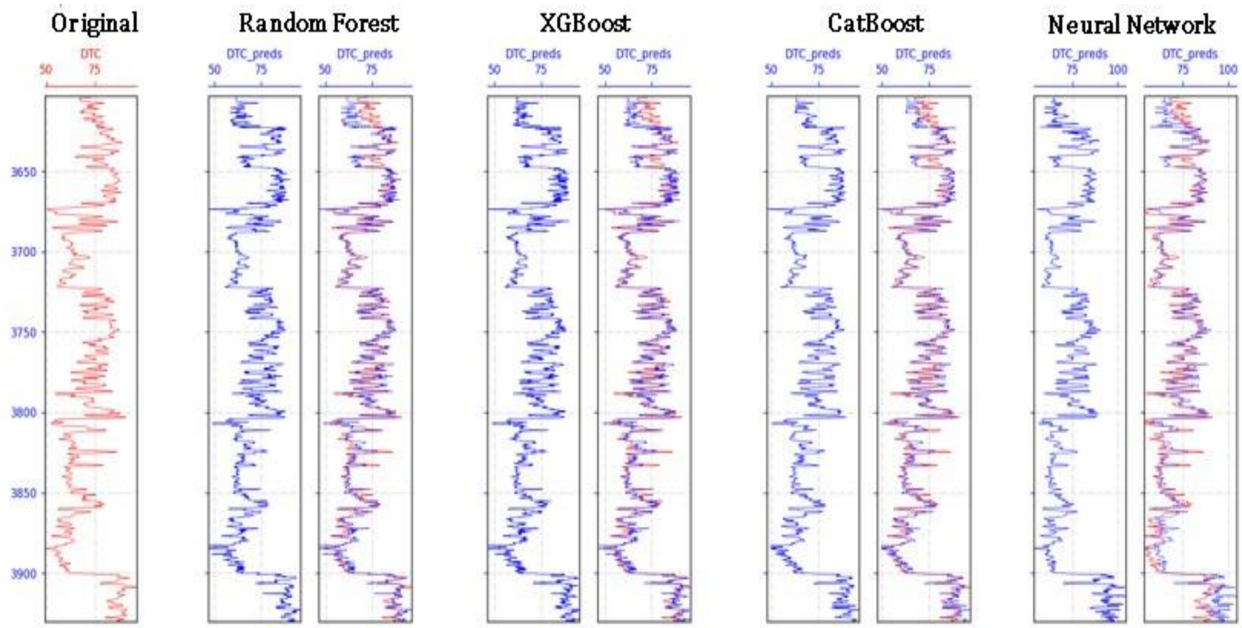

Figure 4. 8: Summary of Blind Well Log Prediction Performance.

2.3 DISCUSSION

- **Analysis and Comparison of Machine Learning Model Performance**

Table 4.5 presents a comparative evaluation of four machine learning algorithms: Random Forest, XGBoost, CatBoost, and Artificial Neural Network (ANN). The models were assessed using three datasets: training, testing, and an independent blind well. Three evaluation metrics were used: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2).

3. Model Performance during Training and Testing (Internal Validation)

During the internal validation stage, the ensemble tree-based models showed better predictive performance than the neural network model.

Random.Forest

Random Forest achieved the highest training accuracy, with an R^2 value of 0.991 and the lowest RMSE of 0.935. However, its performance decreased on the testing dataset, where R^2 dropped to 0.942 and RMSE increased to 2.406. This gap between training and testing results suggests a slight tendency toward overfitting, which is common in deep, unpruned Random Forest models.

XGBoost

XGBoost showed a similar pattern, with strong performance during training (R^2 = 0.987, RMSE = 1.131) and good generalization during testing (R^2 = 0.939, RMSE = 2.486).

CatBoost

CatBoost demonstrated a more balanced behavior. It achieved an R^2 of 0.929 on the training set and 0.920 on the testing set. The small difference between training and testing performance (ΔR^2 = 0.009) indicates excellent generalization ability and highlights the effectiveness of CatBoost's built-in regularization and ordered boosting mechanisms in reducing overfitting.

ANN

The ANN model showed lower performance compared with the tree-based models. It achieved an R^2 value of approximately 0.850 for both training and testing datasets. Although its accuracy was lower, the close agreement between training and testing results indicates stable model behavior.

4. Generalization Ability on the Blind Well (External Validation)

Evaluating the models on a blind well that was not used during training or testing provides the most realistic assessment of model performance in practical field applications, as it represents unseen geological conditions.

CatBoost (Best External Performance)

CatBoost achieved the best performance on the blind well, with the highest R^2 value of 0.884 and the lowest prediction errors (MSE = 12.804 and RMSE = 3.578). Interestingly, its blind well performance was even slightly better than its internal testing performance, indicating a strong ability to capture the general petrophysical trends while remaining less sensitive to noise in the training data.

XGBoost and Random Forest

Both models experienced a noticeable decline in performance when applied to the blind well. For Random Forest, MSE increased from 5.790 during testing to 18.022 on the blind well, while R^2 decreased to 0.836. Similarly, XGBoost achieved an R^2 value of approximately 0.840. These results suggest that although both models performed very well within the training domain, their ability to generalize to unseen geological conditions was weaker than that of CatBoost.

Artificial Neural Network (ANN)

ANN produced the lowest predictive accuracy on the blind well, with an R^2 value of 0.805 and relatively high prediction errors (MSE = 21.475 and RMSE = 4.634).

The results show that gradient boosting models, especially CatBoost and XGBoost, outperform Random Forest in blind well prediction, demonstrating the advantage of sequential residual learning in capturing complex relationships in well-log data. Among all models, CatBoost achieved the best overall performance by providing the highest accuracy and the strongest generalization to unseen data, making it the most suitable choice for field-scale applications. In contrast, the ANN model showed relatively lower predictive performance, which may be due to the limited dataset size and the need for further hyperparameter tuning and architectural optimization, although it still maintained stable behavior across datasets.

- **Feature Importance Analysis for the RF model**

The figure 4.9 shows the relative importance of five key petrophysical logs used as inputs for the developed machine learning models, highlighting each variable's contribution to reducing prediction error and improving model performance. The neutron porosity log (NPHI) is the most influential feature (30–35%) due to its strong sensitivity to pore fluid content and lithology, and its direct physical relationship with compressional sonic travel time, where higher porosity leads to slower acoustic wave propagation. The gamma ray log (GR) ranks second (28–32%), reflecting its importance in distinguishing shale content and lithological variations, which strongly affect elastic and acoustic properties of the formation. The deep resistivity log (RT90) comes third (15–18%), as it provides indirect information about fluid type and saturation, helping the model differentiate between hydrocarbon-bearing and water-bearing zones through its relationship with porosity and Archie's law.

- **Visual Log Matching Formulation**

Based on the results of figure 4. 10, the visual comparison of the predicted logs versus depth provides a strong qualitative validation of the machine learning models, complementing the quantitative error metrics. It allows tracking how well each model reproduces sharp anomalies, peaks, and troughs that reflect geological layering. The Random Forest model performs well in stable intervals but shows a noticeable smoothing effect, where sharp peaks in the original DTC log are flattened due to its averaging nature, limiting its ability to capture extreme variations. XGBoost demonstrates better sensitivity to rapid changes and follows the general trends more accurately, although it introduces slight high-frequency oscillations in heterogeneous zones.

In contrast, CatBoost shows the best visual agreement with the true log, closely matching the original curve throughout the entire depth interval while preserving sharp lithological transitions without adding unnecessary noise, making it the most geologically realistic model. The ANN model, however, produces an overly smoothed response that captures only the general trend with depth while failing to reproduce thin beds and sharp boundaries, resulting in loss of important geological detail. Overall, the visual analysis confirms that gradient boosting methods especially CatBoost because that provide the most reliable and high-fidelity synthetic logs, while ANN suffers from lithological smoothing and limited sensitivity to complex nonlinear variations.

5.CONCLUSION

In conclusion, this study developed a comprehensive workflow for predicting compressional sonic logs (DTC) from conventional well logs using machine learning techniques. The results highlight the critical role of feature engineering, data cleaning, and hyperparameter optimization in improving model accuracy and predictive performance. Although ensemble-based algorithms generally outperformed the other models, their effectiveness was influenced by the specific characteristics of each algorithm, particularly with respect to data preprocessing procedures such as scaling and normalization. Furthermore, the incorporation of blind well testing provided a more rigorous and realistic assessment of model generalization capabilities. Among the evaluated models, Random Forest and CatBoost achieved the best overall performance, demonstrating high predictive accuracy and robustness. These findings suggest that CatBoost can serve as a reliable and cost-effective alternative for generating compressional sonic logs in wells where such measurements are unavailable or incomplete.

General Conclusion

The accurate prediction of compressional sonic logs (DTC) from conventional well-log data represents a valuable advancement for both petrophysical and geophysical applications. This study demonstrated that machine learning techniques can effectively model the complex relationships between sonic velocity and routinely acquired logging measurements, providing a reliable alternative when sonic logs are missing, incomplete, or of poor quality. The results confirmed that data preprocessing, feature engineering, and hyperparameter optimization play critical roles in enhancing prediction accuracy and model robustness.

Among the evaluated algorithms, ensemble learning methods, particularly CatBoost, exhibited the best balance between predictive accuracy and generalization capability, successfully reproducing both the overall trends and detailed variations of the measured sonic logs. The integration of blind well validation further demonstrated the reliability of the developed models under unseen geological conditions.

From a petrophysical perspective, the generated synthetic sonic logs can support porosity estimation, lithology characterization, formation evaluation, and reservoir property assessment. From a geophysical perspective, accurate sonic log prediction provides valuable input for seismic-to-well ties, velocity modeling, time–depth conversion, seismic inversion, and geomechanical analyses. Therefore, the proposed workflow offers a practical and cost-effective solution for enhancing subsurface characterization while reducing the need for expensive logging operations. The successful application of machine learning in this context highlights its potential as a powerful tool for improving decision-making and reducing uncertainty in reservoir exploration and development.

BIBLIOGRAPHY

[1] Ministry of Hydrocarbons and Mines, Algeria, accessed 2026, https://energy.gov.dz.

[2] ALNAFT, Exploration opportunities and basin data: accessed June 9, 2026, http://www.alnaft.dz

[3] Mahmoud, H., 2023, Automatic characterization and quantitative analysis of seismic facies in naturally fractured reservoir: Case study of Amguid Messaoud field, Algeria, Mining Mineral Deposits, 17(3), 42–48.

[4] Mahmoud, H., 2023, Automatic characterization and quantitative analysis of seismic facies in naturally fractured reservoir: Case study of Amguid Messaoud field, Algeria, Mining Mineral Deposits, 17(3), 42–48.

[5] World Energy Council, 2007, N–S and NW–SE sections in the Oued Mya Basin, Algeria.

[6] Mitchell, T. M., 1997, Machine Learning, McGraw-Hill, New York.

[7] Nilsson, N. J., 2009, The Quest for Artificial Intelligence, Cambridge University Press.

[8] Russell, S., and P. Norvig, 2010, Artificial Intelligence: A Modern Approach, Prentice Hall.

[9] Cuddy, S., 2021, The benefits and dangers of using artificial intelligence in petrophysics, Artificial Intelligence in Geosciences, 2, 1–10.

[10] Lee, J., 2026, Roadmap on artificial intelligence and machine learning for smart manufacturing, Diva Portal.

[11] Goodfellow, I., Y. Bengio, and A. Courville, 2016, Deep Learning, MIT Press.

[12] Blessing, E. O., B. Umekwe, and G. Gbulie, 2023, Recent application of machine learning algorithms in petroleum geology: A brief review, International Journal of Research in Science and Innovation, X(IX), 91–99.

[13] Bishop, C. M., 2006, Pattern Recognition and Machine Learning, Springer.

[14] Müller, A. C., and S. Guido, 2016, Introduction to Machine Learning with Python, O'Reilly Media.

[15] Dramsch, J. S., 2020, 70 years of machine learning in geosciences, Advances in Geophysics,

61, 1–97.

[16] Nasrabadi, N. M., 2006, Pattern Recognition and Machine Learning, Springer.

[17] James, G., D. Witten, T. Hastie, and R. Tibshirani, 2021, An Introduction to Statistical Learning, Springer.

[18] Halder, R. K., M. N. Uddin, Md. A. Uddin, S. Aryal, and A. Khraisat, 2024, Enhancing K-nearest neighbor algorithm: A comprehensive review and performance analysis of modifications, Journal of Big Data, 11(1), 113.

[19] DEV Community, 2026, K-nearest neighbor (KNN) algorithms for machine learning, accessed 2026, https://dev.to/victoralando/k-nearest-neighborknn-algorithms-for-machine-learning-2n1b.

[20] Stecanella, B., 2021, Support vector machines (SVM) algorithm explained, accessed 2026, https://monkeylearn.com/blog/introduction-to-support-vector-machines-svm.

[21] Hastie, T., R. Tibshirani, and J. Friedman, 2009, The Elements of Statistical Learning, Springer.

[22] Zhou, Z. H., 2012, Ensemble Methods: Foundations and Algorithms, Chapman and Hall/CRC.

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

[24] GeeksforGeeks, 2026, Machine learning algorithms, accessed 2026, https://www.geeksforGeeks.org/machine-learning/machine-learning-algorithms/.

[25] Goodfellow, I., Y. Bengio, and A. Courville, 2016, Deep Learning, MIT Press.

[26] Scribd, 2007, World Energy Council CD Algeria, accessed 2026, https://www.scribd.com/document/337902269/WEC-2007-CD.

[27] SAS Communities, 2026, Where does GenAI fit within the AI landscape, accessed 2026, https://communities.sas.com/t5/SAS-Communities-Library/Where-does-GenAI-fit-within-the-AI-landscape/ta-p/915000.

[28] ResearchGate, 2026, An overview of supervised machine learning methods, accessed 2026, doi:10.20544/HORIZONS.B.04.1.17.P05.

[29] ResearchGate, 2026, Introduction to supervised learning, accessed 2026, https://www.researchgate.net/publication/389860273_Introduction_to_Supervised_Learning.

[30] Akervoll, I., M. S. Talukdar, S. H. Midtlyng, J. A. Stensen, and O. Torsaeter, 2000, WAG injection experiments with in-situ saturation measurements, SPE Improved Oil Recovery Conference, SPE-59323.

[31] Tiab, D., and E. C. Donaldson, 2024, Petrophysics: Theory and Practice of Measuring Reservoir Rock and Fluid Transport Properties, Elsevier.

[32] Isah, A., et al., 2025, A review of data-driven machine learning applications in reservoir petrophysics, Arabian Journal for Science and Engineering, 50(24), 20343–20377.

[33] Ellis, D. V., and J. M. Singer, 2007, Well Logging for Earth Scientists, Springer.

[34] Asquith, G. B., D. Krygowski, and C. R. Gibson, 2004, Basic Well Log Analysis, AAPG.

[35] Doveton, J. H., 1994, Geologic Log Analysis Using Computer Methods, AAPG.

[36] Rider, M. H., 1996, The Geological Interpretation of Well Logs, Whittles Publishing.

[37] Serra, O., 1984, Fundamentals of Well-Log Interpretation, Elsevier.

[38] Zoback, M. D., 2007, Reservoir Geomechanics, Cambridge University Press.

[39] Bacon, M., R. Simm, and T. Redshaw, 2003, 3-D Seismic Interpretation, Cambridge University Press.

[40] Tiab, D., and E. C. Donaldson, 2015, Petrophysics, Gulf Professional Publishing.

[41] Schlumberger, 1989, Log Interpretation Principles/Applications.

[42] Dewan, J. T., 1983, Essentials of Modern Open-Hole Log Interpretation, PennWell Books.

[43] AAPG, 2004, Basic Well Log Analysis, American Association of Petroleum Geologists.

[44] Weller, J. M., 1960, Stratigraphic Principles and Practice, Harper & Row.

[45] Nelson, P. H., 1994, Permeability–porosity relationships in sedimentary rocks, The Log Analyst, 35(3).

[46] Wyllie, M. R. J., A. R. Gregory, and G. H. F. Gardner, 1956, Elastic wave velocities in heterogeneous and porous media, Geophysics, 21(1), 41–70.

[47] Saggaf, M. M., and E. L. Nebrija, 2003, Estimation of missing logs by regularized neural networks, AAPG Bulletin, 87(8), 1377–1389.

[48] Feng, R., D. Grana, and N. Balling, 2021, Imputation of missing well log data by random forest and uncertainty analysis, Computers & Geosciences, 154, 104789.

[49] Al-Fakih, A., et al., 2024, Well log data generation and imputation using sequence-based generative adversarial networks, arXiv preprint arXiv:2412.00718.

[50] Garini, S. A., et al., 2024, Filling well: An effective technique to handle incomplete well log data for lithology classification, MethodsX, 13, 103127.

[51] Gama, P. H. T., et al., 2025, Imputation in well log data: A benchmark for machine learning methods, Computers & Geosciences, 196, 105789.

[52] Pham, N., X. Wu, and E. Z. Naeini, 2020, Missing well log prediction using convolutional LSTM network, Geophysics, 85(4), WA159–WA171.

[53] Saggaf, M. M., and E. L. Nebrija, 2003, Estimation of missing logs by regularized neural

networks, Journal of Geophysics and Engineering, 4(1), 1–15,

[54] Lin, L., et al., 2023, Missing well log reconstruction using a sequence self-attention deep learning framework, Geophysics, 88(6), D391–D410.

[55] Ren, Q., et al., 2023, Reconstruction of missing well logs using facies-informed wavelet transform, SPE Journal, 28(6), 2946–2963.

[56] Guyon, I., S. Gunn, M. Nikravesh, and L. A. Zadeh, 2006, Feature Extraction: Foundations and Applications, Springer.

[57] Saleh, K., 2026, GeoXMate: A user-friendly GUI software for machine learning applications in petroleum geoscience, Research Square preprint.

[58] Little, R. J. A., and D. B. Rubin, 2019, Statistical Analysis with Missing Data, Wiley.

[59] Hair, J. F., W. C. Black, B. J. Babin, and R. E. Anderson, 2019, Multivariate Data Analysis, Cengage Learning.

[60] Montgomery, D. C., and G. C. Runger, 2018, Applied Statistics and Probability for Engineers, Wiley.

[61] Liu, L., 2008, BEST: Bayesian estimation of species trees under the coalescent model, Bioinformatics, 24(21), 2542–2543.

[62] Aggarwal, C. C., 2017, Outlier Analysis, Springer.

[63] Prokhorenkova, L., et al., 2018, CatBoost: Unbiased boosting with categorical features, NeurIPS.

[64] Chen, T., and C. Guestrin, 2016, XGBoost: A scalable tree boosting system, KDD Proceedings, 785–794.

[65] Myers, J. L., A. D. Well, and R. F. Lorch, 2010, Research Design and Statistical Analysis, Routledge.

[66] Géron, A., 2022, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, O'Reilly Media.

[67] James, G., D. Witten, T. Hastie, and R. Tibshirani, 2021, An Introduction to Statistical Learning (2nd ed.), Springer.

[68] Guyon, I., S. Gunn, M. Nikravesh, and L. A. Zadeh, eds., 2006, Feature extraction: Foundations and applications: Berlin, Springer, 520 p.