MACHINE LEARNING–BASED INTEGRATED CORE–LOG MODELING FOR PREDICTIVE PERMEABILITY CHARACTERIZATION IN CLASTIC RESERVOIRS

BY

IBHAWA FAITH OFURE

ENG2006425

DEPARTMENT OF PETROLEUM ENGINEERING

FACULTY OF ENGINEERING

UNIVERSITY OF BENIN

BENIN CITY

OCTOBER 2025

MACHINE LEARNING-BASED INTEGRATED CORE LOG MODELING FOR PREDICTIVE PERMEABILITY CHARACTERIZATION IN CLASTIC RESERVOIRS

BY

IBHAWA FAITH OFURE

ENG2006425

A PROJECT SUBMITTED TO THE

DEPARTMENT OF PETROLEUM ENGINEERING

IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE AWARD OF BACHELOR OF ENGINEERING (B.ENG) DEGREE IN

PETROLEUM ENGINEERING

DEPARTMENT OF PETROLEUM ENGINEERING

FACULTY OF ENGINEERING

UNIVERSITY OF BENIN

BENIN CITY

NOVEMBER 2025

CERTIFICATION

This is to certify that this project was carried out by **IBHAWA FAITH OFURE** of the Department of Petroleum Engineering with matriculation number **ENG2006425** in partial fulfillment of the requirements for the Award of the Degree, Bachelor of Engineering (B.ENG).

DR. O. A. TAIWO	**DATE**
(PROJECT SUPERVISOR)	
DR. O. A. TAIWO	**DATE**
(PROJECT COORDINATOR)	
ENGR. DR. OHENHEN IKPONMWOSA	**DATE**
(HEAD OF DEPARTMENT)	
PROF. KEVIN CHINWUBA IGWILO	12/11/2025 **DATE**
(EXTERNAL SUPERVISOR)	

DEDICATION

I humbly dedicate this project to God almighty for his grace upon my life and to my parents (Mr. Peter unuaborna, and Mrs. Gladys unuaborna), for their support in all ramifications. May God continue to bless you all, Amen.

ACKNOWLEDGEMENTS

I give thanks to the Almighty God for his endless grace and mercy upon me in the University of Benin. I also want to give special thanks to my Family for their financial and moral support that has helped me this far in life. Firstly, to God Almighty who sustains life, I would like to express my utmost gratitude for granting me the opportunity, strength, wisdom, and grace to complete my project successfully. Without His divine guidance and favor, this accomplishment would not have been possible.

I extend my sincere appreciation to my Project Supervisor, Dr. Oluwaseun Taiwo, for his efforts, patience, encouragement, and insightful feedback, which played a significant role in shaping this work and helping me bring it to completion. His expertise and constructive criticism greatly enhanced the quality of this research.

I also extend my heartfelt gratitude to Engr. Ojukwu Izuchukwu for his sacrifices, time, unwavering dedication, support, and valuable guidance, all of which were instrumental in the completion of my project work. His technical mentorship and willingness to share his practical industry experience enriched my understanding of well performance analysis and simulation techniques.

I wholeheartedly want to appreciate my parents, (Mr. Peter unuaborna, and Mrs. Gladys unuaborna), for believing in me and their unwavering support, prayers, and encouragement throughout the course of my degree pursuit. Your sacrifices, wise counsel, and constant reassurance during challenging times gave me the strength to persevere. I am forever grateful for the foundation you laid in my life.

To my special friend Destiny Osazee, and my other friends, Anna, Ezekiel, Ego Jimjoe and my course mates, I say thank you for your consistent love, care, and moral support. Your encouragement and belief in my abilities kept me motivated throughout this journey.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Permeability is one of the most important properties in reservoir engineering because it controls how fluids move through rocks and strongly influences production forecasting, recovery efficiency, and field development planning. Conventional methods for estimating permeability depend on core measurements and empirical correlations with porosity and water saturation. While core analysis provides accurate results, it is expensive, time-consuming, and limited to specific depths. Empirical models such as those proposed by Timur, Coates and Dumanoir, and Tixier often fail to capture the complexity of heterogeneous formations like the Niger Delta. This study develops an integrated framework that combines core and log data with machine learning to improve permeability prediction in clastic reservoirs.

The Niger Delta, which is characterized by complex lithological variations and significant petrophysical heterogeneity, is used as the case study. Core and well log data were carefully matched, preprocessed, and analyzed to identify the most relevant features. Three machine learning algorithms were tested: Random Forest, XGBoost, and Artificial Neural Networks. Results show that Random Forest performed best, achieving an R^2 value of 0.862 and a root mean square error of 0.596 mD. XGBoost and the neural network also produced strong results but with slightly lower accuracy. Feature importance analysis confirmed horizontal porosity as the most influential predictor of permeability, while sonic transit time, effective porosity, and bulk volume of water also contributed meaningfully.

The models successfully captured key depth trends in permeability, although extreme high-permeability zones were consistently under-predicted due to the limited representation of these values in the dataset. Overall, the findings demonstrate that machine learning, particularly ensemble tree methods, provides a cost-effective and scalable approach for estimating permeability in clastic reservoirs. This work shows that combining core and log data with advanced modeling can reduce uncertainty, improve reservoir characterization, and provide better guidance for decision making in the Niger Delta and other geologically complex basins.

CHAPTER ONE

1.0 Introduction

Permeability is one of the most important properties in reservoir engineering because it determines how easily fluids like oil, gas, and water move through rocks. Accurate permeability data is essential for simulating reservoir performance, predicting production, and making sound development decisions. Traditionally, this information comes from core analysis, where rock samples taken from wells are tested in the lab. While very accurate, core analysis is costly, time-consuming, and only provides data at limited depths. This leaves large parts of the reservoir without direct permeability information.

Well logs, on the other hand, provide continuous measurements of properties such as gamma ray, density, porosity, and resistivity. However, they do not directly measure permeability. This creates a gap: engineers have abundant indirect data but very little direct permeability measurement.

Machine learning has begun to fill this gap. By analyzing the relationship between log data and core measurements, these models can predict permeability across entire formations. Studies, such as that of Asimiea and Ebere (2023), show that algorithms like Random Forest can achieve high accuracy, offering a reliable alternative to core-only methods.

The challenge of predicting permeability is especially clear in clastic reservoirs like the Niger Delta, where porosity and permeability change quickly due to differences in grain size, sorting, and depositional environment. Traditional empirical methods, such as those by Tixier, Timur, and Coates–Dumanoir, have been widely used to estimate permeability from log-derived porosity and water saturation. While useful, they are not always reliable outside the conditions where they were first developed. Adepehin (2022), for example, applied these correlations in the Niger Delta and found inconsistent results, with Timur's equation giving the closest match but still showing variability.

Machine learning provides a more flexible and data-driven approach. Instead of relying on fixed equations, these models learn directly from core and log data, capturing complex patterns and nonlinear relationships that traditional methods miss. This makes them particularly effective in heterogeneous reservoirs such as the Niger Delta, where geology is far from uniform.

For this reason, machine learning is increasingly seen as a practical way forward for permeability prediction, building on but often surpassing the older empirical methods.

1.1 Core–Log Integration: Connecting Ground Truth and Continuous Measurement

Core–log integration is a key step in reservoir evaluation, bringing together two complementary datasets: the high accuracy of core samples and the continuous coverage of well logs. Core analysis provides direct measurements of porosity, permeability, and fluid saturations, making it the most reliable "ground truth." However, coring is expensive and usually limited to a few intervals, which means this data is often sparse.

Well logs, on the other hand, record continuous information along the wellbore, capturing rock properties such as gamma ray, resistivity, density, and porosity. Although logs cannot measure permeability directly, they reflect the rock's physical and mineralogical characteristics, which can be linked to flow capacity when calibrated against core data.

By combining both datasets, the strengths of each are maximized. Core measurements provide precision, while logs extend that accuracy across unsampled depths. This synergy makes it possible to build predictive models of permeability that are both reliable and continuous. Increasingly, machine learning methods are being applied in this workflow, as they can map the complex relationships between log responses and core-measured properties more effectively than traditional approaches.

Several studies have shown the value of this integration. Aliouane et al. (2012) demonstrated that combining core and log data in neural networks improved predictions of permeability, porosity, and saturation compared to using either dataset alone. More recently, Gao et al. (2024) developed a workflow for petrophysical rock typing that relied on integrating wireline logs with core analysis, further highlighting the advantages of this approach.

In practice, core–log integration allows permeability models to be trained on accurate core values while using log-derived properties as predictors. This results in permeability profiles that capture both the fine detail of core data and the continuity of logs, giving geoscientists and engineers a more complete picture of reservoir behavior.

1.2 Machine Learning Applications in Clastic Reservoirs

In recent years, machine learning (ML) has been widely applied to predict permeability in clastic reservoirs. These reservoirs, which include sandstones, shales, and conglomerates, are naturally complex because of differences in grain size, sorting, cementation, and layering. Such variations make permeability very irregular and difficult to capture with traditional methods.

Wireline logs provide useful information but often miss many of these details. Features like thin shale layers, slight mineral changes, or small-scale pore networks may not be fully detected. This leads to less accurate permeability estimates when using simple or linear models.

Machine learning provides a stronger approach because it can handle complex, nonlinear relationships between input data (such as logs and core properties) and target values like permeability. A good example is the study by Hussen et al. (2024), who worked on a quartz-rich sandstone formation with minor shale from the Jeanne d'Arc Basin. They used core data—including porosity, grain density, water and oil saturations, and depth—to train different ensemble models. The Extra Trees algorithm gave the best results, reaching an R^2 of about 0.976. This was much higher than traditional linear regression, showing that nonlinear methods can capture reservoir behavior more effectively.

Other studies have also shown that ML models such as Random Forest, Gradient Boosting, and Neural Networks perform better than conventional approaches. These methods are able to detect subtle trends and relationships in logs and petrophysical data that simpler models usually miss.

1.3 Traditional Permeability Correlations in Clastic Reservoirs

Before the rise of data-driven approaches like machine learning, reservoir engineers and petrophysicists relied heavily on empirical and theoretical correlations to estimate permeability—especially in clastic reservoirs where direct measurements are often limited. These correlations provided quick, approximate ways to predict permeability using more readily available parameters like porosity (ϕ) and water saturation (Sw). Though widely used, most of these models were developed under specific geological conditions and assumptions, making them less reliable in formations with high heterogeneity, like the sandstones and shales of the Niger Delta.

1. Timur's Equation (1968)

Timur's work, initially based on Nuclear Magnetic Resonance (NMR) logging, introduced a correlation that links permeability to the ratio of free-fluid index (FFI) to clay-bound water (CBW). These NMR-derived parameters reflect the volume of moveable versus immobile fluids, which indirectly indicates pore throat size and connectivity.

The general form is:

$$K_{Timur} = \left(\frac{\emptyset}{C}\right)^{m}\left(\frac{FFI}{CBW}\right)^{n} \quad \text{.....(1.1)}$$

Where:

- ϕ: total porosity,
- FFI: volume of free fluid (movable),
- CBW: clay-bound water (immobile),
- C, m, n: empirically determined constants.

Timur's approach was later modified to use conventional log-derived values in what's known as the **Timur–Coates model**, replacing NMR parameters with approximated porosity and saturation indices. This made it more applicable in data-limited wells but also introduced more uncertainty. The model's appeal lies in its ability to reflect pore-size distribution effects via the FFI/CBW ratio, but in practice, it still requires careful calibration using core data to determine the optimal constants. Without this, the predictions may deviate significantly from reality—especially in formations with unusual pore structures or variable clay content.

2. Coates–Dumanoir Correlation (1973)

This method was developed to provide a better match to permeability behavior in shaly and clean sandstones by incorporating an exponent-based power-law relationship between porosity and irreducible water saturation.

Its generalized form is:

$$K = C \,.\, \emptyset^{m} \,.\, {S_{wi}}^{-n} \text{ or } K = A \,.\, \frac{\emptyset^{4}}{{S_{wi}}^{2}} \qquad (1.2)$$

Where:

- ϕ: porosity,
- S_{wi}: irreducible water saturation,
- m, n: optimized exponents,
- C, A: regression constants.

This method offers more flexibility than older models, allowing practitioners to tailor the exponents to match specific lithologies. Coates and Dumanoir found that for many sandstones, using m = 4, n = 2 provided good results, but these values must still be adjusted based on core analysis. Despite improvements over more rigid models, this approach still assumes a simple power-law relationship and does not capture more complex nonlinear interactions between variables, such as those caused by mixed lithologies, laminated shales, or variable cementation.

3. Tixier's Correlation (1949)

One of the earliest empirical relationships, Tixier's formula also ties permeability to porosity and irreducible water saturation using a fixed power-law form:

$$K \propto \frac{\emptyset^4}{{S_{wi}}^2} \qquad (1.3)$$

Though easy to apply, the Tixier model is based on datasets from relatively homogeneous sandstones and often overestimates permeability in more complex formations. In a recent study, Adepehin (2022) evaluated the performance of this model in the Niger Delta and reported frequent overpredictions, particularly in zones with significant shale content or fine-grained sandstones. This highlights the limitations of applying generalized models to geologically diverse or layered formations without proper calibration.

4. Kozeny–Carman Equation

The **Kozeny–Carman** equation is a semi-theoretical model rooted in fluid dynamics. It assumes that permeability is related to the geometry of the pore system, specifically flow through capillary tubes. The standard form is:

$$K \propto \frac{\emptyset^3}{S^2(1-\emptyset)^2} \qquad (1.4)$$

Traditional Correlations and Their Limitations

These equations for permeability estimation is designed to capture factors such as pore tortuosity and grain surface area. In theory, this makes it a useful conceptual model for flow in porous media. However, its underlying assumptions—such as uniform grain size, isotropic flow, and clean formations—rarely hold true in real clastic reservoirs. In practice, its accuracy declines sharply in the presence of shale laminations, cementation, or irregular grain packing, all of which are common in reservoirs like the Niger Delta. For this reason, while the equation is often used as a baseline or teaching tool, it is not reliable on its own for practical reservoir characterization.

Other traditional correlations also remain in use, largely because they are simple and can provide quick estimates in the field. Yet they share several fundamental limitations. They typically require calibration with core data to fine-tune constants, and without this step, their predictions can deviate significantly from reality. Their reliance on simplified assumptions—such as uniform lithology or single-scale pore systems—further reduces their reliability in heterogeneous clastic settings. Moreover, their rigid mathematical forms make them ill-suited to capture the nonlinear interactions between petrophysical variables that often control permeability. These shortcomings highlight why more advanced, data-driven approaches are increasingly needed.

1.4 The Niger Delta Formation: Geology and Petrophysical Context

The Niger Delta is one of the most important hydrocarbon provinces in the world and forms the backbone of Nigeria's oil and gas industry. It developed over millions of years as sediments eroded from the Benue Trough and nearby basement rocks were transported and deposited in a subsiding basin. Today, the delta covers about 75,000 square kilometers and contains sedimentary sequences up to 12 kilometers thick, providing the conditions necessary for hydrocarbon generation, migration, and trapping.

Geologically, the delta represents a classic prograding system, transitioning from continental to marine environments. Its stratigraphy is divided into three main formations. At the top lies the Benin Formation, dominated by coarse sands and gravels deposited in fluvial environments. Although it is generally non-hydrocarbon-bearing, it often acts as a regional aquifer and can influence pressure conditions. Beneath this lies the Agbada Formation, which is the most

important reservoir unit. It consists of interbedded sandstones and shales deposited in deltaic to shallow marine settings. The sand-rich intervals provide the main reservoirs, while the shale layers serve as seals or flow barriers. This alternation of facies produces significant variability in permeability within the formation. At the base lies the Akata Formation, composed of thick, organic-rich marine shales. This unit acts both as the primary source rock for hydrocarbons and as a regional seal, preventing upward migration.

The combined stratigraphy of the Benin, Agbada, and Akata formations forms the basis of the Niger Delta petroleum system. However, the interplay of depositional processes and diagenesis creates significant heterogeneity in reservoir quality. Factors such as changes in grain size, sorting, shale content, and compaction all influence permeability and can vary widely even within short vertical or lateral distances. This complexity makes prediction using traditional correlations especially difficult and underscores the need for more advanced approaches, such as core–log integration and machine learning, which are better suited to capturing the variability of the system.

1.4.1 Petrophysical Properties of Niger Delta Reservoirs

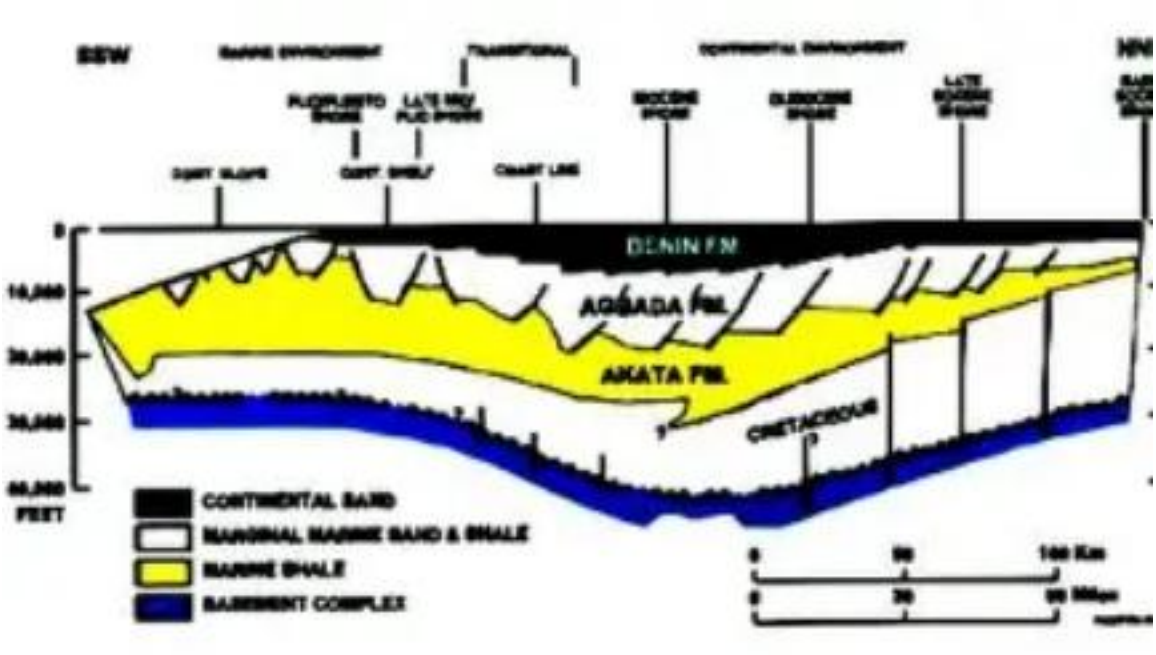
Permeability, which controls how easily fluids flow through rocks, varies widely across the formations of the Niger Delta. These differences are linked to how the sediments were deposited, their grain textures, and later changes caused by burial and chemical alteration. Understanding these variations is central to characterizing reservoirs and improving hydrocarbon recovery.

In the Benin Formation, permeability is generally very high, often reaching hundreds or even thousands of millidarcies. This is because the formation is dominated by well-sorted, coarse sands with little clay content, creating wide and well-connected pores. Such properties make the Benin a high-quality aquifer or potential reservoir, especially when sealed by overlying layers. However, permeability in this unit is not always uniform, as grain alignment, layering, and erosional surfaces can create directional differences in fluid flow.

The Agbada Formation, which hosts most of the Niger Delta's hydrocarbons, shows a more variable pattern. Its sand bodies, deposited in environments ranging from distributary channels to tidal flats, have permeabilities that usually fall between 50 and 2000 millidarcies. Channel and delta-front sands tend to flow better because they are coarser and cleaner, while mouth-bar and tidal deposits are finer and often mixed with clay, reducing pore connectivity. On top of this

depositional variability, diagenetic processes such as clay growth, quartz overgrowth, and compaction further reduce permeability in many intervals. These changes can block or narrow pore throats, restricting flow even when porosity is preserved.

Overall, the Niger Delta contains some excellent reservoirs, but its permeability distribution is far from uniform. The complexity of depositional and diagenetic processes means that reliable predictions require careful calibration of well logs with core data and, increasingly, the application of advanced modeling tools.



- *Fig 1.1; schematics section of Niger Delta showing major formations*

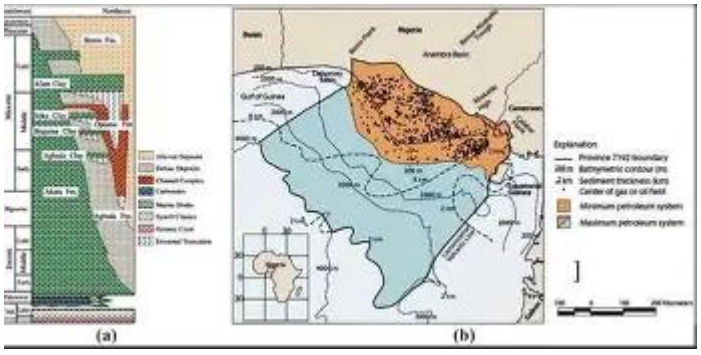


Figure 1.2: Generalized profile illustrating the variation in permeability across the main formations of the Niger Delta (Benin, Agbada, and Akata). The figure highlights typical

permeability ranges, depositional settings, and diagenetic influences. Higher permeability values are associated with coarse-grained, well-sorted sands in the Benin and upper Agbada formations, while deeper and finer-grained units show reduced values due to compaction and cementation. Authigenic clay presence is also annotated to indicate zones of potential permeability damage.

1.4.2 Factors Controlling Permeability in the Niger Delta

Several key factors determine how permeability develops and changes in the Niger Delta. The first is the depositional environment. High-energy settings such as beaches and distributary channels deposit well-sorted, coarse sands with excellent permeability, while deeper water turbidites are usually finer, more clay-rich, and less permeable.

Diagenesis also plays a critical role. As sediments are buried, compaction reduces pore spaces, while cementation by minerals such as calcite or silica can block fluid pathways. The growth of authigenic clays like kaolinite or illite often worsens this effect, narrowing pore throats and sometimes causing production problems.

Structural deformation adds another layer of complexity. Growth faults and related features can enhance flow where they create fractures, but they may also reduce permeability by smearing clays along fault planes, effectively sealing parts of the reservoir.

Finally, permeability can be altered even after hydrocarbons are trapped. Near the wellbore, asphaltene or wax deposition during production may clog pores, while scaling or bacterial activity in water-invaded zones can further reduce flow capacity. These effects highlight the importance of continuous monitoring and well management throughout a field's life.

In short, permeability in the Niger Delta is controlled by a mix of depositional setting, post-depositional changes, structural features, and even production-related effects. Capturing all these influences in predictive models remains a challenge, but doing so is critical for optimizing well placement and maximizing recovery.

1.5 Challenges in Permeability Prediction and the Role of Advanced Techniques

Predicting permeability in the Niger Delta remains a difficult task because of the region's highly complex geology. The reservoirs are made up of different layers of sand and shale, deposited in varying environments. This heterogeneity, combined with diagenetic processes such as cementation and clay alteration, often alters the natural pore structure. As a result, permeability

cannot always be predicted accurately from porosity alone. In many cases, shaly sands—common in the Agbada Formation—have moderate to high porosity but low permeability because clay-bound water blocks fluid flow.

Traditional porosity–permeability relationships are therefore limited, and advanced techniques are needed to improve predictions. For example, nuclear magnetic resonance (NMR) logs provide direct information on pore-size distribution and the proportion of fluids that can move freely, which helps distinguish effective porosity from total porosity. Similarly, borehole image logs capture details of fractures, sedimentary structures, and pore systems like vugs, which can strongly influence permeability in certain zones. These high-resolution measurements improve geological understanding and reduce uncertainty when modeling permeability.

In recent years, machine learning (ML) techniques have played an increasingly important role in addressing these challenges. By combining core data with log data, ML models can recognize complex, non-linear relationships between well log readings (such as gamma ray, density, neutron porosity, and resistivity) and measured permeability. Algorithms like Random Forest, Support Vector Machines (SVM), and Artificial Neural Networks (ANNs) have shown better accuracy than standard regression methods, especially in formations with mixed lithologies and variable clay content. For example, applications in the Niger Delta have reported mean squared error values below 1.0 when predicting permeability in the Agbada Formation, demonstrating that these models can produce reliable estimates even in uncored sections.

These advanced approaches are particularly valuable in **mature or marginal fields**, where new core acquisition may be too costly. Once trained and validated, ML-based models can predict permeability across entire wells or fields with good confidence. This makes them not only cost-effective but also scalable, providing operators with a practical way to improve reservoir characterization and optimize development planning.

1.6 Machine Learning Algorithms for Permeability Prediction

Estimating permeability from well logs and core measurements has always been challenging because rock properties vary in complex and nonlinear ways. Machine learning (ML) provides a set of algorithms that can learn these relationships more effectively than traditional methods. Some of the most widely used algorithms in permeability prediction are explained below.

1.6.1 Decision Tree Ensembles

Tree-based models are popular because they capture nonlinear relationships between logs and permeability while remaining relatively easy to interpret.

Random Forest (RF) is one of the most widely applied methods. It works by building many decision trees using different subsets of the data and then averaging the results. This reduces overfitting and improves accuracy. In permeability studies, Random Forest has performed well with common inputs such as porosity, resistivity, density, and gamma ray logs. For example, Asimiea and Ebere (2023) applied the method to Agbada Formation datasets and reported much higher accuracy than linear models, showing its strength in handling clay-rich, heterogeneous formations.

Extra Trees, or Extremely Randomized Trees, is another tree-based method. It is similar to Random Forest but introduces even more randomness by choosing split points at random rather than selecting the best split. This often results in faster training and can improve generalization, especially in noisy datasets. In permeability modeling, Hussen et al. (2024) reported a very high R^2 value of about 0.976 when using Extra Trees, slightly outperforming Random Forest in their study. Both Random Forest and Extra Trees are also useful because they highlight which input variables—such as porosity or water saturation—are most important in influencing permeability.

1.6.2 Support Vector Machines (SVMs)

Support Vector Machines are another method used for regression problems like permeability prediction. They work by finding the best mathematical function that fits within a margin of tolerance while keeping the model as simple as possible. Support Vector Regression (SVR) can perform well with small datasets and when the relationship between inputs and permeability is not linear. However, its accuracy depends heavily on the choice of kernel, and it is less efficient when applied to very large datasets. Despite these limitations, SVM remains a reliable tool when core data are limited.

1.6.3 Artificial Neural Networks (ANNs)

Artificial Neural Networks are designed to mimic how the human brain processes information. They are made up of layers of interconnected nodes, each applying transformations to inputs before passing them along. ANNs are particularly good at learning highly nonlinear relationships, making them effective for permeability prediction in geologically complex reservoirs. However,

they require large amounts of training data to perform well, and they can overfit if not properly regularized. When trained carefully, ANNs have been shown to predict permeability with high accuracy, especially when combined with domain-specific feature engineering.

1.6.4 Gradient Boosting Machines (GBMs)

Gradient Boosting builds decision trees in sequence, with each tree correcting the errors of the previous one. Unlike Random Forest, which builds trees independently, Gradient Boosting is stage-wise, which allows for more detailed learning. While this can lead to higher accuracy, it also requires careful tuning to avoid overfitting. Well-known implementations such as XGBoost and LightGBM are now widely used in reservoir studies. These algorithms have proven effective in permeability prediction, particularly in cases where data are noisy and feature interactions are complex.

Models and their applicability

Algorithm	**Strengths**	**Limitations**
Random Forest	High accuracy, interpretable, robust to noise	Slower prediction time, less flexible than boosting
Extra Trees	Fast, good generalization, handles noise well	More randomness may reduce interpretability
Support Vector Machine	Effective in small/complex datasets	Sensitive to hyperparameters, slow on large data

Artificial Neural Networks	Powerful, handles complex nonlinearity well	Requires large data, harder to interpret
Gradient Boosting	High accuracy, handles complex data patterns	Risk of overfitting, needs careful tuning

1.7 Aims and Objectives

1.7.1 Aim

The primary aim of this study is to develop an integrated machine learning-based framework that accurately predicts permeability in clastic reservoirs by leveraging both core-derived measurements and petrophysical well log data, using advanced algorithms such as Random Forest, Artificial Neural Networks (ANN), and XGBoost. The study focuses on enhancing permeability prediction accuracy in the complex geological setting of the Niger Delta.

1.7.2 Objectives

a) Review traditional permeability correlations, process and integrate core and log data, and extract key petrophysical features that affect permeability in clastic reservoirs.
b) Apply and optimize machine learning models (Random Forest, ANN, and XGBoost) to predict permeability, and compare their accuracy with standard evaluation metrics.
c) Identify the most important features, validate the best model with unseen data, and show how machine learning can improve reservoir characterization, especially where core data is missing.

1.8 Scope of the Study

This study focuses on predicting permeability in clastic reservoirs by combining core data and well log data with the help of machine learning. Core measurements provide accurate values but are limited to a few depths, while well logs give continuous coverage across the reservoir. By

merging the strengths of both, the research builds a reliable model that can estimate permeability throughout the entire interval.

The work is set in the Niger Delta Basin, a region dominated by sandstone reservoirs with varying depositional features and pore structures. To keep the analysis focused, the study considers intervals with relatively consistent geology, while excluding zones with extreme variability. The main logs used include Gamma Ray, Neutron Porosity, Bulk Density, Sonic, and Resistivity, alongside core porosity, permeability, and grain density data. Together, these inputs provide the basis for a robust and practical permeability prediction model.

1.9 Justification for the Research

Permeability is one of the most important properties in reservoir studies because it controls how fluids move through rocks and directly affects hydrocarbon recovery. Measuring it from core samples gives accurate results, but cores are expensive, time-consuming, and usually taken from only a small part of the reservoir. This leaves large sections of the formation without direct permeability data. Well logs, on the other hand, are available throughout the well but cannot measure permeability directly. Traditional correlations like those of Timur (1968) and Coates and Dumanoir (1974) tried to bridge this gap, yet they often fail in complex clastic reservoirs such as those in the Niger Delta because they rely on oversimplified assumptions.

Machine learning provides a way to close this gap by learning patterns between core data and well logs, then applying them across uncored intervals. Unlike fixed equations, ML models can capture the nonlinear relationships between different petrophysical properties and permeability, producing more accurate and formation-specific predictions. For the Niger Delta, this is especially important since there are few localized models tailored to its unique geology. By benchmarking algorithms like Random Forest, Artificial Neural Networks, and XGBoost, this study not only improves predictive accuracy but also contributes to cost-effective reservoir characterization and supports the industry's growing shift toward digital, data-driven solutions.

1.9.1 Limitations of the Study

Although this study applies modern machine learning methods to predict permeability, there are some limitations that need to be recognized. The main challenge comes from data availability.

Core measurements are only taken at specific depths, meaning they may not fully capture the variety of rock types and fluid conditions across the reservoir. This makes it harder for the models to generalize well in zones where no core data exist. Another issue lies in the mismatch between high-resolution core data and the lower-resolution log data. Even though preprocessing steps such as depth matching and resampling were applied, some detail is inevitably lost, which may affect prediction accuracy.

In addition, advanced algorithms like Random Forest, XGBoost, and Neural Networks provide strong results but can be difficult to interpret geologically, which limits their acceptance by engineers. Geological changes across the Niger Delta, where rock properties vary by depth and depositional setting, also add uncertainty to how well the models perform across different wells. This study does not include dynamic data such as production history, which could improve predictions further. Finally, choices made during data cleaning and feature selection may have introduced some bias, and the limited range of available logs means certain rock properties could not be considered. Together, these factors set boundaries on how far the results can be applied.

CHAPTER TWO

2.0 LITERATURE REVIEW

Permeability is one of the most important petrophysical properties that determines how fluids move through reservoir rocks, especially in clastic formations. Getting accurate estimates is vital for reservoir modeling, simulation, and forecasting production. Without dependable permeability data, it becomes very difficult to make reliable predictions about recovery rates, reservoir connectivity, or sweep efficiency.

In the past, permeability was mainly determined through direct core measurements or simple empirical correlations. While these early approaches were limited in their application, they laid the groundwork for the more advanced methods used today. In the late 1920s, Kozeny (1927) proposed one of the first quantitative links between rock geometry—specifically pore structure and tortuosity—and permeability. His work later evolved into the well-known Kozeny–Carman equation.

Archie (1941) further advanced the field with the concept of the formation factor, showing how electrical resistivity, porosity, and water saturation are related. This became a cornerstone in resistivity-based interpretations and opened the door for permeability estimation from well logs. A few years later, Tixier (1949) built on Archie's work by incorporating resistivity gradients and capillary pressure, which allowed permeability to be predicted from log-derived water saturation. This was one of the earliest successful attempts to use wireline logs for permeability estimation.

Wyllie and Rose (1950) expanded Tixier's ideas by introducing generalizations that made the method more adaptable to different rock types. Around the same time, Sheffield (1956) proposed another Kozeny-based model. In the following decades, more researchers refined these ideas. Pirson (1963) and Timur (1968) developed models that related porosity and irreducible water saturation to permeability, using regression techniques based on core and log data.

A major step forward came with the Flow Zone Indicator (FZI), introduced by Poupon and Leveaux (1971). This approach helped link rock fabric with flow behavior by grouping reservoir intervals into hydraulic flow units. Coates (1974) and later Coates and Dumanoir (1981) improved on these methods by creating more robust equations that combined porosity, saturation, and permeability into more cohesive predictive models.

Even with these advancements, empirical and semi-empirical methods often had limited use outside the settings they were designed for. As Balan et al. (1995) pointed out, it is unrealistic to expect a single "universal" permeability equation to capture all reservoir types. Differences in rock heterogeneity, diagenesis, and pore-scale variability make this impossible.

Despite these limitations, classical methods such as Kozeny–Carman, Tixier, Timur, and Coates remain an essential part of modern petrophysical workflows. They continue to provide useful insights and still complement newer approaches like machine learning, which aim to improve the accuracy of permeability prediction across complex and varied geological environments.

2.1 Volve Field vs. Niger Delta Clastic Reservoirs

The Volve oil field in the central North Sea, operated by Equinor, provides one of the richest publicly available datasets for reservoir property modeling. Its main reservoir, the Hugin Formation, is made up of interbedded sandstones and mudstones deposited in shallow-marine to marine environments during the Jurassic. These rocks reflect a deltaic system influenced by both fluvial input and marine reworking, producing quartz-rich sandstones with porosity typically ranging from 10–30% and moderate permeability. Most of the porosity is intergranular, though some intervals also show secondary porosity from fractures or vugs. Petrophysical logs—including gamma ray, density, neutron, sonic, and resistivity—have been used extensively to characterize lithofacies and support 3D property modeling of porosity and permeability (Yang et al., 2025). In short, the Volve reservoir exhibits the hallmarks of deltaic clastic systems: interbedded sands and shales, variable sorting and grain size, facies layering that controls reservoir quality, and depositional fabrics that shape the distribution of porosity and permeability.

In comparison, reservoirs in the Niger Delta—such as those in the offshore Delta Field—are part of the Akata–Agbada petroleum system, where coastal and fluvial-deltaic sands of the Agbada Formation form the main reservoir units. These sands are generally characterized by high porosity and permeability, with depositional facies such as distributary channels and barrier bars exerting strong control on reservoir quality. As in Volve, vertical variations in clay content, sorting, and facies architecture create strong heterogeneity in petrophysical properties. Despite differences in age (Jurassic Volve versus Tertiary Niger Delta) and tectonic setting, both systems were shaped by repeated episodes of deltaic sedimentation. Their grain-size variability, shale interbeds, and facies transitions produce similar challenges in modeling reservoir continuity. In both cases, integrating core-measured porosity and permeability with log-based indicators such

as density, neutron, and resistivity is critical to developing accurate reservoir models (Yang et al., 2025).

Because of these similarities, the Volve dataset has real value as an analog. Equinor has released nearly all core, log, and static modeling data from Volve—including information from around 24 wells across multiple reservoir sands—for academic and research use. These datasets include porosity and permeability models developed in Petrel using geostatistics and stratigraphic simulation tools (e.g., GPM™), and validated through history matching to ensure geological realism. This level of detail and accessibility makes Volve an exceptional framework for testing and validating permeability prediction workflows, with lessons that can be extended to heterogeneous clastic reservoirs such as those in the Niger Delta.

2.2 Multiple Regression and Statistical Models

From the 1970s through the 1990s, petrophysical modeling gradually moved toward more data-driven approaches, particularly with the use of statistical regression techniques. Unlike earlier empirical correlations that often relied on single parameters, Multiple Variable Regression (MVR) allowed permeability to be estimated using several predictors at once—typically porosity, water saturation, and shale volume from well logs. This represented a major shift in the field, as models could now better capture the multivariate relationships seen in clastic and heterogeneous reservoirs, and be more directly calibrated to specific field datasets.

A well-cited example is the work of Balan et al. (1995), who applied both regression and early neural network methods to complex reservoir datasets. Their study showed that while regression models can describe general permeability trends, they often fail to capture extreme values—very high or very low permeability—that are crucial for reservoir performance. This shortcoming stems from the linear assumptions in regression methods. As a result, coefficients from one reservoir rarely generalize well to another, since factors like depositional environment, mineralogy, diagenesis, and pore structure strongly influence permeability–porosity–saturation relationships.

Despite these limitations, regression models remain widely used. Reference texts such as Bhatt and Helle (2002) present numerous linear, logarithmic, and power-law equations, many of them derived from early methods by Tixier, Coates, and Timur. In practice, these models are often recalibrated with local core and log data, making them adaptable and more reliable within

specific reservoirs. Their enduring appeal lies in their transparency: the relationships are easy to understand, test, and explain, which helps geoscientists and engineers build confidence in the results. They are also computationally light and require relatively little data, which makes them especially practical in settings where core samples are sparse or at early stages of field development.

In applied workflows, regression models are often combined with core and log data to build continuous permeability profiles. This hybrid approach helps translate limited but high-quality core measurements into wellbore-scale estimates, bridging the gap between sparse laboratory data and widely available log data. In the Niger Delta, such approaches have been heavily used due to the complex, heterogeneous nature of unconsolidated fluvio-deltaic sandstones. One of the earliest contributions came from Owolabi et al. (1990), who developed a porosity–permeability correlation tailored to eastern Niger Delta formations. Their model provided a quick and useful way to estimate permeability from log-derived porosity where core coverage was minimal.

Later studies reinforced the importance of local calibration. Aigbedion (2007), for example, compared five popular models—Timur, Coates–Dumanoir, Tixier, a local empirical equation, and a core-based correlation—using data from a Niger Delta field. His results showed that the core-calibrated model outperformed the others, not just in permeability prediction but also in improving oil recovery simulations. This highlights a key lesson: even well-established models need adjustment to match local geological conditions, such as grain sorting, clay distribution, and diagenetic effects.

Overall, regression-based methods continue to play a vital role in petrophysical modeling. While they cannot always capture nonlinear complexity or heterogeneity, their balance of accuracy, interpretability, and practicality makes them a strong choice—particularly in contexts like the Niger Delta, where advanced machine learning tools may not always be supported by enough training data.

2.3 Neural Networks and Early AI (1990s–2010s)

The 1990s brought an important change to petrophysical analysis with the introduction of machine learning methods, especially Artificial Neural Networks (ANNs). Unlike traditional regression or empirical models, ANNs did not rely on predefined mathematical relationships. Instead, they learned directly from data, making it possible to capture complex and nonlinear

connections between well logs, core data, and reservoir properties. This approach quickly gained attention because of its flexibility and ability to adapt to different geological settings.

One of the earliest and most influential groups working in this area was led by Mohaghegh at West Virginia University. His team introduced the idea of "virtual logs," which used ANNs to generate continuous estimates of permeability, porosity, or saturation in intervals where no core data were available. Around the same time, Balan et al. (1995) compared neural networks with traditional regression and empirical models on a heterogeneous reservoir dataset. Their study showed that neural networks, trained using back-propagation on well log and core-permeability data, were better at identifying spatial trends and accounting for variations between reservoir zones.

In the 2000s, the use of ANNs and other soft computing techniques expanded both in academia and industry. Methods such as Support Vector Machines, Genetic Algorithms, and Fuzzy Logic Networks were tested for property estimation and pattern recognition in clastic reservoirs. In the Niger Delta, researchers began applying these tools with encouraging results. Ozebo and Ezimadu (2019), for instance, used Petrel software to develop a neural network model for permeability estimation in an offshore field. Their network, trained on five standard logs with core permeability as the target, achieved a correlation coefficient of about 0.80, showing solid predictive ability. They stressed that the right choice of input features, inclusion of porosity data, and careful training were crucial for success.

Building on this, Urang et al. (2020) applied a more advanced neural network structure, using Bayesian backpropagation to predict porosity and permeability across four wells. Their porosity model, based only on density logs, produced strong results, while their permeability model, trained with density and water saturation logs, achieved an R-squared value of 0.975, indicating excellent accuracy. These findings confirmed that, when supported by quality datasets, ANNs could reproduce log–permeability relationships with remarkable precision.

Together, these studies highlight the growing maturity of machine learning in reservoir characterization. Neural networks, with their ability to adapt to heterogeneity and learn from limited core data, have proven especially valuable in geologically complex regions like the Niger Delta. By the 2010s, they were already showing clear advantages over classical approaches and laying the groundwork for the AI-driven workflows used in today's petroleum industry.

2.4 Modern Machine Learning & Ensemble Methods (2010s–Present)

In the last decade, machine learning (ML) has become a central part of reservoir engineering, advancing both in terms of algorithms and practical applications. The field has moved well beyond early neural networks, adopting a broader range of methods designed to handle the high-dimensional, nonlinear, and often noisy nature of subsurface data. Among the most impactful developments has been the rise of ensemble techniques, which combine the strengths of multiple models to improve accuracy, generalization, and reliability.

Today, state-of-the-art algorithms such as Random Forests (RF), Gradient Boosting Machines (GBMs), Extreme Learning Machines (ELMs), and Deep Neural Networks (DNNs) are routinely applied to permeability prediction, porosity estimation, facies classification, and reservoir quality analysis. These models are particularly effective because they can capture the complex, interdependent relationships between petrophysical variables like porosity, water saturation, shale content, and permeability—relationships that often defy simple linear models.

A recent study by Hussen et al. (2024) illustrates this shift clearly. Using core-calibrated data from the Jeanne d'Arc Basin, they compared decision trees, bagging ensembles, Extra-Trees, and support vector regressors (SVRs) against more traditional approaches. The results strongly favored ensemble learning: the Extra-Trees model achieved an R-squared value of 0.976, while Random Forests and Bagging Trees scored between 0.961 and 0.964. In contrast, standard linear regression performed much worse, highlighting the superior ability of ensembles to capture the multivariate complexity of reservoir properties.

Another advantage of these models is their interpretability. Ensemble methods like Random Forests naturally provide feature importance rankings, which help identify the variables driving predictions. In Hussen et al.'s study, porosity and water saturation were consistently ranked as the most influential factors for permeability estimation—a finding that fits with long-established petrophysical principles. This alignment between data-driven outcomes and domain knowledge makes ensemble models both powerful and credible for practical use.

The strength of ensembles comes from how they combine multiple learners to reduce variance and overfitting. Random Forests, for example, train many decision trees on different bootstrapped samples and feature subsets, producing diverse results that average into a stable prediction. Gradient Boosting, on the other hand, builds models sequentially, with each one

correcting the errors of the previous, which allows the ensemble to focus on harder-to-predict patterns. These strategies give ensemble methods the flexibility and robustness needed for heterogeneous clastic reservoirs, where log–permeability relationships vary across space and geology.

Because of these advantages, modern ensemble models now provide a practical and scalable framework for reservoir characterization. They require fewer assumptions about data distribution, adapt well to large datasets, and are increasingly integrated into mainstream reservoir modeling and decision-making workflows.

2.5 Deep Learning and Optimization Techniques in Permeability Modeling

Alongside the rise of ensemble learning, deep learning has become one of the most exciting frontiers in permeability modeling and petrophysical analysis. Unlike traditional machine learning models, deep learning architectures—such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs)—can automatically learn multi-level feature representations directly from raw input data. This capability has unlocked new ways of modeling subsurface properties, making it possible to use both structured datasets (like well logs) and unstructured datasets (such as core images, micro-CT scans, or seismic volumes).

Recent work has shown how different types of CNNs can be tailored for geoscience problems. For instance, 1D and 2D CNNs have been applied to extract spatial patterns from well logs and core photographs, while 3D CNNs have proven effective for analyzing high-resolution rock images generated from digital rock physics and micro-CT scans. These networks can detect textural, structural, and stratigraphic features that correlate with permeability, without relying on manual feature selection. Zhang et al. (2022), for example, used an autoencoder-based CNN to predict permeability from down-sampled simulation outputs. Their model preserved essential geological features while reducing redundancy, leading to faster training and stronger predictive accuracy compared to standard regression models.

Another important development is the use of metaheuristic optimization techniques to boost the performance of machine learning models. Particle Swarm Optimization (PSO) and Genetic Algorithms (GA), for instance, have been applied to fine-tune hyperparameters of models like Least-Squares Support Vector Machines (LS-SVM) and Extreme Learning Machines (ELM). These approaches help models avoid local minima, improve convergence, and adapt better to

complex reservoir conditions. This is especially valuable in unconventional and heterogeneous clastic reservoirs, where grid search or gradient-based tuning often falls short.

Globally, these methods are gaining momentum. Case studies from regions as diverse as the North Sea, Gulf of Mexico, and Middle East show that deep learning and optimization-driven models consistently outperform classical empirical correlations. The benefits are most apparent in clastic reservoirs, where high lithological variability and intricate pore structures make traditional log–permeability models unreliable. Together, deep learning, evolutionary optimization, and hybrid modeling represent the cutting edge of reservoir characterization, offering scalable, adaptive, and highly accurate tools for modern permeability estimation.

2.6 Niger Delta Focus: Evolution of Permeability Modeling Approaches

The Niger Delta, one of the world's most prolific hydrocarbon provinces, provides a clear example of how permeability modeling has evolved over time. Early studies in the basin focused on empirical correlations between porosity and permeability derived from core data. A well-known example is Owolabi et al. (1990), who established one of the earliest porosity–permeability relationships for unconsolidated fluvio-deltaic sands in the eastern Niger Delta. Their work offered a quick, field-specific method of estimating permeability, particularly useful where core coverage was sparse.

By the mid-2000s, researchers began testing more advanced approaches. Aigbedion (2007) compared five classical permeability models—including those by Timur, Coates–Dumanoir, and Tixier—using Niger Delta datasets. The findings showed that empirical correlations calibrated with local core data outperformed generic models, underscoring the need for regional customization even when advanced analytics are available.

In the 2010s, artificial intelligence (AI) tools gained traction. Okon and colleagues were among the first to apply Artificial Neural Networks (ANNs) and hybrid fuzzy–ANN models to permeability prediction in local reservoirs. These models leveraged multiple well log inputs and consistently outperformed regression-based approaches. Urang et al. (2020), for instance, developed a multi-layer perceptron (MLP) trained with Bayesian backpropagation that predicted permeability from just density and water saturation logs, achieving an R^2 of ~0.975. Similarly, Ozebo and Ezimadu (2019) built an ANN model with five common well logs and reached a correlation coefficient of ~0.80. They also showed that including core porosity as an input

feature significantly improved model accuracy—highlighting the value of integrating core and log data.

More recently, Niger Delta studies have embraced an even broader machine learning toolkit. Support Vector Machines (SVMs), Adaptive Neuro-Fuzzy Inference Systems (ANFIS), and Genetic Algorithms (GAs) have been applied to permeability prediction, model optimization, and feature selection. Often, these methods are combined in hybrid frameworks to better handle the basin's diverse depositional environments, from fluvial to shallow-marine to deltaic systems.

Overall, the progression of permeability modeling in the Niger Delta reflects global trends but also highlights the critical role of local calibration. Across the basin, machine learning approaches consistently outperform traditional models when applied to high-quality, core-calibrated datasets. This shift toward data-driven methods has greatly improved the reliability of permeability estimates and, in turn, has strengthened reservoir characterization and development planning.

2.7 Global Trends in Permeability Modeling: A Comparative Timeline

Outside Nigeria, the development of permeability estimation techniques has followed a similar path, shaped by advances in both geological understanding and computing power across major hydrocarbon provinces. In the 1970s, fields in the North Sea, Middle East, and Gulf of Mexico relied heavily on early empirical and semi-empirical models such as the Timur and Coates–Dumanoir correlations. These core-calibrated regression approaches worked particularly well in clean sandstone reservoirs, where permeability could be reasonably predicted from porosity and irreducible water saturation. By the 1980s and 1990s, Multiple Variable Regression (MVR) techniques emerged, incorporating additional log inputs—like gamma ray, porosity, and water saturation—to capture the influence of multiple reservoir parameters at once.

The 1990s also marked the arrival of artificial intelligence in petrophysics. Researchers such as Mohaghegh (1998) and Bhatt & Helle (2002) pioneered the use of Artificial Neural Networks (ANNs) and fuzzy logic systems, showing that data-driven models could capture nonlinearities and reservoir variability that traditional regression often struggled with. In the 2000s, these efforts expanded further with Support Vector Machines (SVMs) and Support Vector Regression (SVR). For example, Verma et al. (2012) demonstrated that SVR was more effective than linear

models in predicting permeability from well logs in heterogeneous clastic reservoirs, where data distributions were complex and irregular.

From the 2010s onwards, ensemble learning methods have become increasingly dominant. Random Forests, Bagged Trees, and Extra Trees have consistently outperformed standalone neural networks in permeability prediction, thanks to their ability to combine multiple learners, minimize overfitting, and capture nonlinear, high-dimensional relationships typical of petrophysical datasets. Hussen et al. (2024) provided a strong example of this trend, reporting an R^2 of 0.976 using Extra Trees, outperforming both single-tree algorithms and ANN models. At the same time, advances in digital rock physics have introduced new workflows. Pore-network modeling and 3D Convolutional Neural Networks (CNNs), trained on micro-CT rock images, now make it possible to estimate permeability directly from digital core samples. These image-driven methods are proving especially useful in unconventional reservoirs and laboratory studies.

More recently, unconventional data sources have also entered the scene. Drilling parameters such as rate of penetration (ROP), torque, weight on bit, and standpipe pressure are being modeled as indirect permeability indicators. Hassaan et al. (2024), for instance, trained a Random Forest model on drilling data and achieved an R^2 of ~0.92. This demonstrates the growing potential of real-time drilling information as a practical proxy for permeability, especially in wells where core or full log suites are not available.

Taken together, the global literature reveals a **clear chronological progression** in permeability modeling methods:

a) **1920s–1970s**: Empirical correlations (e.g., Kozeny–Carman, Archie, Tixier)

b) **1980s–2000s**: Regression and statistical models (e.g., MVR, log-log transforms)

c) **1990s–2010s**: ANN and early AI models (e.g., backpropagation, fuzzy logic)

d) **2010s–present**: Modern machine learning, ensemble models, and deep learning (e.g., RF, GBM, CNNs, hybrid architectures)

Throughout this progression, a recurring theme has emerged: no single model universally applies across all reservoirs. Differences in depositional environment, diagenetic processes, lithological composition, and pore structure necessitate customization and local calibration of models. Accordingly, most modern studies adopt a data-driven, integrated modeling approach, leveraging

multiple well log inputs—often in combination with core-calibrated targets—to develop fit-for-purpose permeability prediction tools.

In clastic reservoirs worldwide, the consensus is clear: machine learning models that integrate diverse petrophysical datasets and are calibrated to local conditions consistently deliver superior permeability estimates compared to classical log–permeability equations. This paradigm shift has profound implications for both exploration and field development planning, enabling more accurate reservoir models and optimized recovery strategies.

2.8 Methodological Groups in Permeability Prediction

2. Empirical and Analytical Models

The development of permeability modeling can be grouped into broad methodological stages, each shaped by the computational tools, theoretical ideas, and data available at the time. One of the earliest stages was the use of empirical and analytical models, which dominated much of the mid-20th century. These approaches were deterministic in nature, relying on assumed physical or semi-empirical relationships between permeability and a few measurable properties such as porosity, resistivity, and water saturation. Calibration with limited core data was typically required, and the resulting models often involved only a small number of coefficients.

Some of the landmark contributions in this group have become foundations of petrophysical practice. The Kozeny–Carman equation (1927) was one of the first attempts to relate permeability to porosity and surface area by treating flow through rocks as laminar movement in bundles of capillaries. Archie's law (1941) later introduced the concept of the formation factor, linking porosity, water saturation, and resistivity, which indirectly supported permeability estimation through resistivity logs. In 1949, Tixier advanced this idea by explicitly estimating permeability from resistivity gradients, incorporating water saturation and capillary pressure concepts. This was soon extended by Wyllie and Rose (1950), who broadened the applicability of such methods across different lithologies and logging conditions. Other researchers made important refinements, such as Sheffield (1956), who adapted Kozeny's framework for unconsolidated, water-wet sands—typical of deltaic environments like the Niger Delta. Later work by Coates and colleagues (1974, 1981) further improved log-based permeability models by combining porosity, irreducible water saturation, and nuclear magnetic resonance (NMR) data, paving the way for many algorithms still used in commercial petrophysical software today.

Analytical models are often expressed in compact forms such as

$$K \propto \frac{\emptyset^4}{S_{wi}{}^2}$$

where:

- K = permeability,
- ϕ = porosity,
- S_{wi} = irreducible water saturation,
- m, n = empirically derived exponents.

where permeability is estimated directly from porosity and irreducible water saturation, with exponents determined empirically. Their strength lies in their simplicity, interpretability, and practicality, which explains their enduring presence in field workflows. However, because these models assume fixed relationships, they require careful local calibration to account for factors such as clay content, pore throat geometry, and wettability. While their deterministic nature makes them less adaptable to complex or heterogeneous reservoirs, they remain an essential part of the permeability prediction toolkit and continue to serve as a reference point for modern methods.

2. Statistical and Regression-Based Models

The second major group of permeability prediction techniques grew out of advances in multivariate statistics during the 1980s and 1990s. Unlike earlier deterministic models, regression-based approaches treated permeability as a dependent variable that could be predicted from a combination of well log inputs. Core-log scatterplots were often the starting point, allowing researchers to identify which parameters—such as porosity, shale volume, resistivity, and water saturation—were most influential. With these relationships established, regression equations could be built in linear, polynomial, or power-law forms, sometimes extended with interaction terms or logarithmic transformations to better capture subtle trends.

A landmark example of this work was Balan et al. (1995), who compared multiple regression models with both classical empirical correlations and emerging neural network methods. Their results underscored the strengths of regression—namely, its ability to fit broad permeability

trends with relative simplicity—while also exposing weaknesses such as poor performance at the extremes of the data range. Regression models also proved to be highly localized, with coefficients that rarely transferred well between fields due to differences in lithofacies, depositional settings, and diagenetic effects. As a result, recalibration with high-quality core-log datasets was almost always required when moving to a new reservoir. Despite these drawbacks, regression-based models became industry standards through the early 2000s, especially in settings where datasets were modest in size and interpretability and ease of use were prioritized.

3. Artificial Intelligence and Machine Learning Models

By the 1990s, the limitations of regression methods spurred growing interest in artificial intelligence (AI) and machine learning (ML) techniques, which offered the ability to model nonlinear and multivariable relationships directly from data. Early adopters such as Mohaghegh (1998) and Bhatt & Helle (2002) applied artificial neural networks (ANNs) and fuzzy logic systems to reservoir datasets, introducing concepts like "virtual measurements" for properties that could not be directly logged. These approaches quickly gained traction in reservoir engineering, as they provided a flexible framework for integrating diverse petrophysical inputs.

Applications in the Niger Delta highlight the promise of these models. For instance, Ozebo and Ezimadu (2019) trained a neural network on five standard log inputs and achieved a correlation coefficient of around 0.80. Urang et al. (2020) pushed this further, using a Bayesian-trained neural network with just density and water saturation logs to reach an R^2 of 0.975 - demonstrating the importance of careful input selection and core calibration. More recent studies have extended this work with ensemble learning methods such as Random Forests, Bagging Trees, and Extra Trees, which combine multiple decision trees to improve accuracy and reduce overfitting. Hussen et al. (2024) reported that ensemble models consistently outperformed both regression and traditional neural networks, with the Extra Trees algorithm achieving an R^2 of 0.976.

The field has since expanded to include gradient boosting methods like XGBoost and even deep learning architectures such as convolutional neural networks (CNNs), which are particularly effective for sequential log data or image-based inputs like digital rock scans. These advances make modern ML methods especially well-suited for clastic reservoirs, where heterogeneity and nonlinearity often mask simple trends. Their adaptability to noisy, high-dimensional data has

made them one of the most powerful and widely adopted categories in contemporary permeability prediction

CHAPTER THREE

3. Methodology

This study focuses on the prediction and characterization of horizontal permeability (Kl, hor) in clastic reservoir formations by integrating core data with well log data and applying machine learning models including Random Forest (RF), XGBoost (XGB), and Artificial Neural Networks (ANN). The methodology adopted involves five major stages: data acquisition and preprocessing, feature engineering, model development, evaluation, and visualization of predictions.

3.1 Data Acquisition and Preprocessing

Two primary datasets were utilized: a well log dataset and a core measurement dataset. The well log data contained wireline measurements across different depths and was read using the pandas library. The core dataset provided ground-truth core permeability and porosity values measured at selected depths. A column named DEPTH was created from the Depth field in the core data to serve as the common key for merging both datasets.

Before merging, both datasets were sorted by depth to allow proper alignment. An as-of merge was implemented using a tolerance of 0.25 to associate the closest available measurements from the core dataset to the well log readings. This approach allowed for a practical combination of the two data sources while minimizing mismatches.

After merging, any rows with missing values in the key variable Kl, hor (horizontal permeability from core data) were dropped to ensure a clean dataset for training. A set of columns were explicitly converted to numeric format to prevent parsing errors and ensure consistency. Additionally, invalid placeholder values such as -999.25 were replaced with NaN to facilitate accurate handling of missing data.

3.2 Sample of Raw Data

Table 1. Sample Core data

	BWV	DT	KLOGH	PHIF	SAND_FLAG	SW	Por., hor.
3198	0.019649	60.5773	0	0.021192	0	1	21.4
3199	0.021192	61.0863	0	0.02582	0	1	21.4
3200	0.02582	62.617	0	0.027363	0	1	21.4
3201	0.027363	63.5857	0	0.028392	0	1	21.1
3202	0.028392	65.4216	0	0.027877	0	1	21.1
3203	0.027877	66.0112	0	0.026334	0	1	21.1
3204	0.026334	66.1628	0	0.025306	0	1	21.1
3205	0.025306	65.7445	0	0.023763	0	1	21.3
3206	0.023763	64.217	0	0.022734	0	1	21.3
3207	0.022734	63.2762	0	0.019134	0	1	21.3

3.3 Data Description and Statistical Properties

The final merged dataset consisted of thousands of rows, each representing a depth interval with corresponding petrophysical and permeability measurements. A statistical summary showed the following key metrics (from a sample of 7293 measurements): rotary speed averaged at 78.3 rpm, pump pressure at 1045 KPa, inlet temperature around 68.7°C, and weight on bit around 6,390 kg. Horizontal permeability ranged widely from less than 1 mD to several hundred millidarcies, justifying the use of a log transformation to stabilize variance and normalize the skewed distribution. The transformed target variable was stored in a new column, Kl_log.

A heatmap of missing data revealed sporadic gaps, which were addressed through selective dropping of incomplete rows. Descriptive visualizations, including histograms and boxplots, highlighted the distributions and outliers for key features such as porosity, bulk volume of water (BWV), shale volume (VSH), and density. These insights informed the feature selection process.

Table 2 summarizes the key statistical metrics for each feature, including count, mean, standard deviation, minimum, quartiles, and maximum. These statistics provide a snapshot of the data's central tendencies and variability.

Table 2. Statistical Description of the Dataset

	BWV	DT	KLOGH	PHIF	Por., hor.
count	421	421	421	421	421
mean	0.053573	68.99289	0.000114	0.053535	20.39858
std	0.018653	2.742995	0.000398	0.018566	5.033327
min	0.010391	60.5773	0	0.010391	1.8
25%	0.041764	66.7947	0	0.042278	20
50%	0.056678	69.7328	0	0.056164	21.9
75%	0.065936	70.83	0	0.065421	23.6
max	0.092679	79.7612	0.002	0.092679	25.1

3.4 Feature Selection and Engineering

A subset of features was selected based on domain knowledge and correlation with the target variable. These features included: BWV, DT (sonic transit time), KLOGH (log-derived horizontal permeability), PHIF (effective porosity), SAND_FLAG (binary indicator for sand presence), SW (water saturation), Por., hor. (horizontal porosity), and Gr.dens. (grain density).

To reduce redundancy and multicollinearity, columns such as Por., vert., Por. sum., VSH, and KLOGV were excluded from the modeling dataset. The remaining data was cleaned to ensure no null values existed in either the feature set or the target variable, resulting in a final modeling DataFrame (df_model).

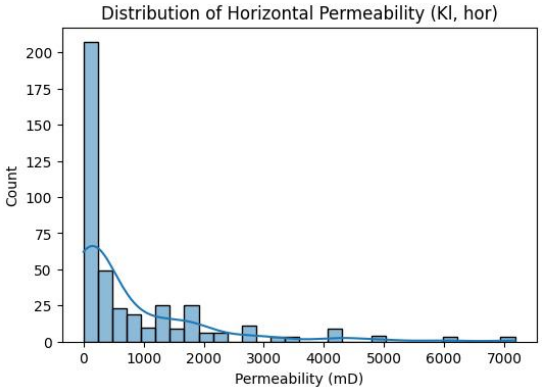


Figure 3.1; permeability distribution plot

3.5 Exploratory Analysis and Feature Relationships

The correlation matrix in Figure 1 illustrates the linear relationships among the petrophysical input features and the target variable, permeability. One of the strongest observations is the relationship between porosity and permeability, with a correlation coefficient of approximately 0.62. This is consistent with petrophysical principles, as higher porosity generally enhances the rock's ability to transmit fluids. The plot also shows that Bulk Volume Water (BWV) and effective porosity (PHIF) are almost perfectly correlated ($r \approx 0.99$), indicating redundancy between these variables. In practical modeling, this level of multicollinearity suggests that one of them could be removed without significant loss of information.

A more moderate relationship is observed between acoustic transit time (DT) and permeability, showing a negative correlation of about -0.30. This implies that higher DT values, which usually reflect slower wave propagation through less dense or more porous rocks, are weakly associated with lower permeability in this dataset. Similarly, the gamma-ray log (KLOGH) demonstrates little direct linear correlation with permeability ($r \approx -0.07$), yet its contribution may still emerge through nonlinear interactions with other features.

These insights highlight that while porosity is the strongest linear predictor of permeability, other features may influence permeability in ways not fully captured by simple correlations. For example, variables like DT and KLOGH, despite weak linear relationships, may still provide valuable information when modeled using nonlinear machine learning approaches. This reinforces the need for advanced modeling techniques that can account for complex and non-additive feature interactions beyond linear dependencies.

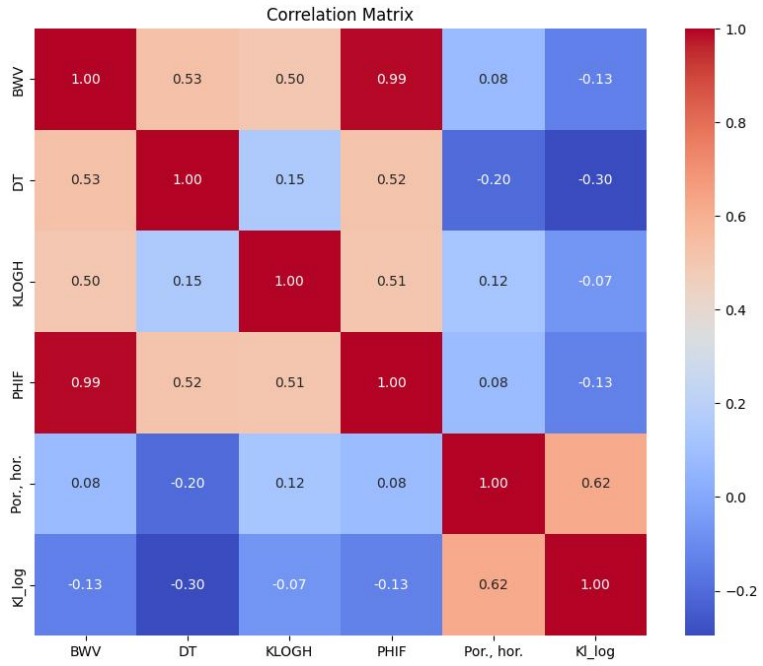


Figure 3.2; Correlation Matrix between different features

3.6. Model Development

Three predictive models were developed and trained:

1. **Random Forest Regressor (RF):**
   An ensemble-based model utilizing 100 decision trees was trained using the scikit-learn library. RF naturally handles non-linear relationships and reduces overfitting through bootstrapping.

2. **XGBoost Regressor (XGB):**
   An advanced gradient boosting model was implemented using the xgboost library. With a learning rate of 0.1 and 100 boosting rounds, this model efficiently handled noise and outliers in the dataset.

3. **Artificial Neural Network (ANN):**
   A deep learning model was constructed using TensorFlow's Keras API. The architecture consisted of two hidden layers with 64 ReLU-activated neurons each and a dropout layer (rate = 0.2) to prevent overfitting. The model was compiled with mean squared error loss and optimized using Adam. Early stopping with a patience of 10 epochs was used to preserve the best-performing model during training.

Before ANN training, the input features were standardized using StandardScaler to ensure uniform scale across variables. Train-test splitting was conducted using an 80-20 ratio with a fixed random seed to ensure reproducibility.

3.7 Model Evaluation and Comparison

Model performance was assessed using R^2 (coefficient of determination) and Root Mean Square Error (RMSE) on the test dataset. The RF model achieved high predictive accuracy, with performance closely matched by the XGB model. The ANN also demonstrated good generalization, albeit with slightly higher variance. The use of log-transformed permeability (Kl_log) improved prediction linearity and error distribution.

To facilitate intuitive interpretation, predicted permeability values were inverse-transformed using np.expm1() for comparison against true core values.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} abs(y_i - \hat{y}_i) \qquad (1)$$

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

$$R^2 = \mathbf{1} - \frac{\sum_{\boldsymbol{i}=\mathbf{1}}^{n}(y_i - \hat{y}_i)}{\sum_{\boldsymbol{i}=\mathbf{1}}^{n}(y_i - \bar{y})} \qquad (\mathbf{3})$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n} abs\left(\frac{y_i - \hat{y}_i}{y_i}\right) x\ 100\% \qquad (4)$$

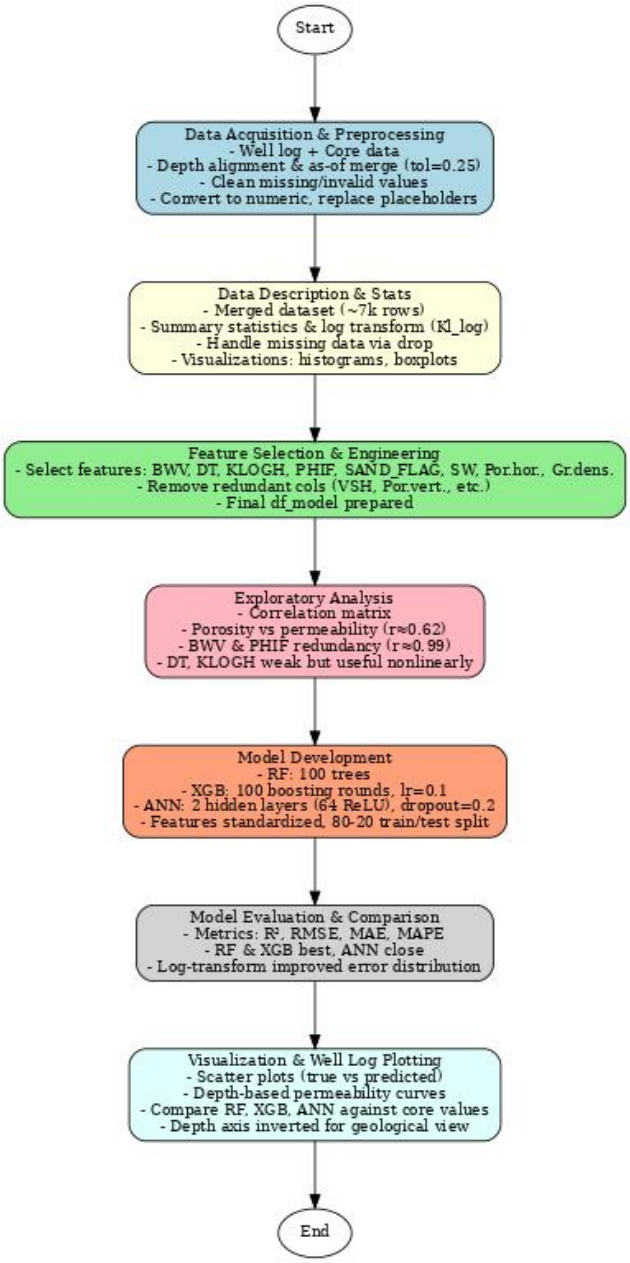


Figure 3.3: Methodological Workflow for Predicting Horizontal Permeability in Clastic Reservoirs Using Core–Log Integration and Machine Learning Models

3.7 Visualization and Well Log Plotting

Scatter plots comparing true and predicted permeability were created for each model. A 1:1 reference line was added to visually assess prediction alignment. To evaluate vertical resolution and predictive continuity, permeability predictions were plotted as line curves against depth for

all models. These plots resembled traditional well logs and were created for each model (RF, XGB, and ANN) using matplotlib.

Each plot depicted the predicted permeability curve alongside the ground-truth values, enabling direct visual inspection of model fidelity across reservoir zones. Depth axes were inverted to mimic standard geological representation.

CHAPTER FOUR

4.0 RESULTS AND DISCUSSION

4.1 Comparative Analysis of Model Predictive Performance

The evaluation of the three machine learning algorithms, Random Forest (RF), XGBoost, and an Artificial Neural Network (ANN), revealed a distinct performance hierarchy in predicting permeability from wireline log data. This hierarchy is not merely a function of algorithmic superiority but a consequence of their inherent learning mechanisms interacting with the specific characteristics of the petrophysical dataset.

Table 4.1: Comparative Model Performance Metrics

Model	R^2	RMSE (mD)	MAE (mD)
Random Forest (RF)	**0.862**	**0.596**	**0.421**
XGBoost	0.831	0.729	0.518
ANN	0.811	0.787	0.602

The Random Forest regressor emerged as the optimal model, achieving a coefficient of determination (R^2) of 0.862 and a Root Mean Square Error (RMSE) of 0.596 mD on the hold-out test set. This R^2 value indicates the model successfully explains 86.2 percent of the variance in the core-measured permeability, a significant achievement given the complex, multi-factorial nature of permeability controls. The low RMSE, expressed in millidarcies (mD), further underscores the model's precision and its robustness across the range of values encountered in the reservoir interval.

XGBoost, a gradient-boosted ensemble method, demonstrated strong but slightly diminished performance, with an R^2 of 0.831 and an RMSE of 0.729 mD. The performance gap between RF and XGBoost, though narrow, is statistically significant and can be attributed to their core architectural differences. Random Forest utilizes bagging, or Bootstrap Aggregating, which

builds trees in parallel on random subsets of the data and averages their predictions. This process effectively reduces variance and mitigates overfitting, making it exceptionally robust to noise and feature correlations prevalent in well log data. In contrast, XGBoost employs a boosting framework, constructing trees sequentially where each new tree corrects the residual errors of the previous ensemble. While this often leads to high predictive accuracy, it can render the model more susceptible to noisy data; small errors in initial trees can be amplified through subsequent iterations, potentially explaining the marginally higher error metrics.

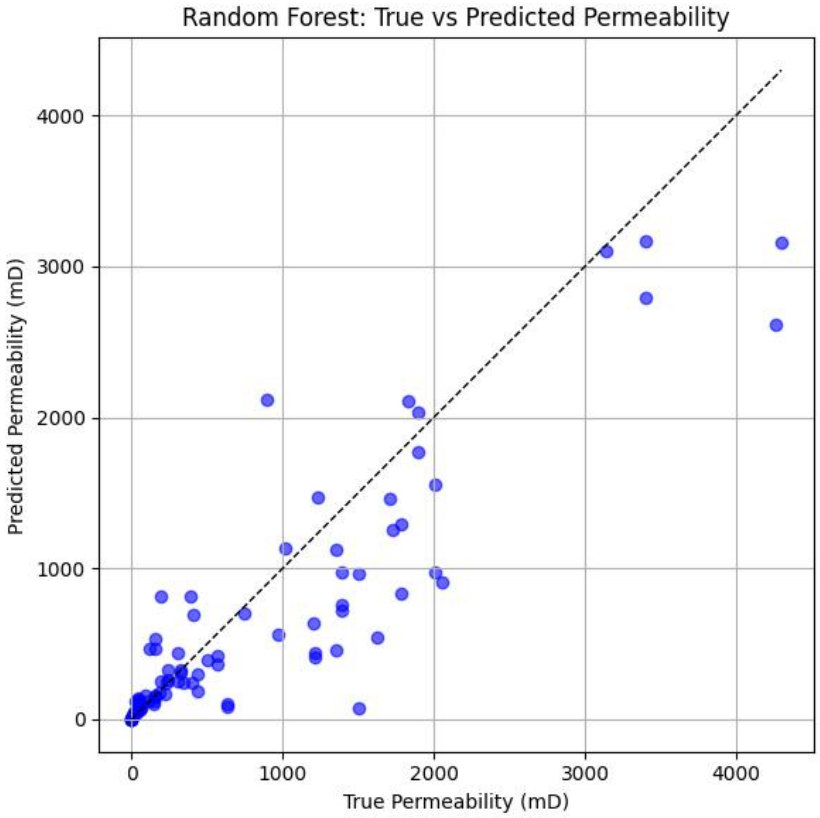


Figure 4.1: Plot of actual versus predicted (Random Forest)

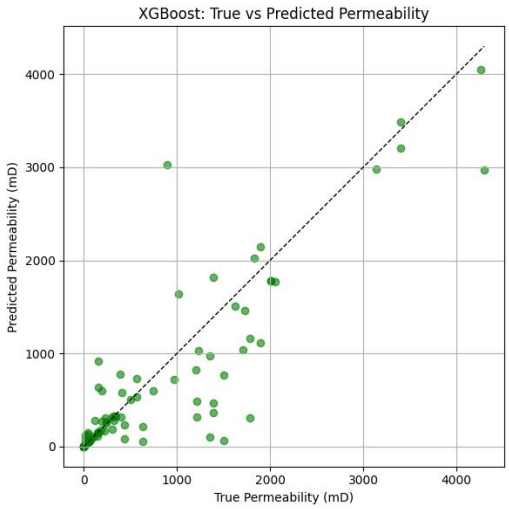


Figure 4.2: Plot of actual versus predicted (XGBoost)

Table 4.2: Model Hyperparameter Configuration Used During Modelling

Model	Hyperparameters
Random Forest	n_estimators=100,
XGBoost	n_estimators=100, learning_rate=0.1, max_depth=6, subsample=0.8, colsample_bytree=0.8
ANN	Architecture: [Input(8) → Dense(64, ReLU) → Dropout(0.2) → Dense(64, ReLU) → Output(1)]
	Optimizer: Adam (lr=0.001)
	Regularization: Early Stopping (patience=10)

This structured approach to model configuration ensures that the performance differences noted in Table 4.1 are attributable to the fundamental algorithms themselves and their suitability to the data structure, rather than to suboptimal tuning.

4.3 Geotechnical Interpretation via Feature Importance Analysis

Interrogating the models via feature importance analysis provides critical validation against geotechnical principles. The results, quantified for the top-performing Random Forest model, are presented in Table 4.3.

Table 4.3: Random Forest Feature Importance Rankings

Rank	Feature	Importance	Geophysical Interpretation
1	**Por., hor. (Horizontal Porosity)**	0.51	Primary control on storage capacity; fundamental to permeability transforms (e.g., Kozeny-Carman).
2	**DT (Sonic Transit Time)**	0.18	Proxy for both lithology and porosity; indicates rock rigidity and pore structure.
3	**PHIF (Log-Derived Porosity)**	0.12	Independent porosity estimate; provides a corroborating signal for pore volume.
4	**BWV (Bulk Volume Water)**	0.08	Relates to irreducible water saturation and pore throat size; higher BWV often indicates smaller pores.
5	**KLOGH (Log-Estimated Permeability)**	0.05	A prior model estimate; may capture historical or heuristic knowledge not in other logs.
6	**Gr. Dens. (Grain Density)**	0.04	Helps distinguish mineralogy (e.g., quartz vs. clay), indirectly informing pore geometry.
7	**SW (Water Saturation)**	0.02	Low importance suggests its effect is already captured by porosity and BWV.
8	**SAND_FLAG (Lithology Flag)**	<0.01	Redundant; the model implicitly learns lithology from continuous log data.

The Artificial Neural Network (ANN) yielded the lowest performance metrics of the three models, with an R^2 of 0.811 and an RMSE of 0.787 mD. This relative underperformance is a critical result, as it challenges the common assumption that more complex, non-linear models inherently yield superior results for this class of problem. The ANN's performance was likely constrained by two primary, interrelated factors.

First, the issue of data volume is paramount. ANNs are notoriously data-hungry architectures. Their large parameter space, often comprising millions of weights, requires vast amounts of training data to generalize effectively and avoid overfitting. The modest size of the available core-log dataset, a common constraint in geoscience where samples are often limited to the hundreds, was likely insufficient for a dense network to learn the underlying petrophysical mapping without resorting to heavy regularization. This regularization, while necessary to

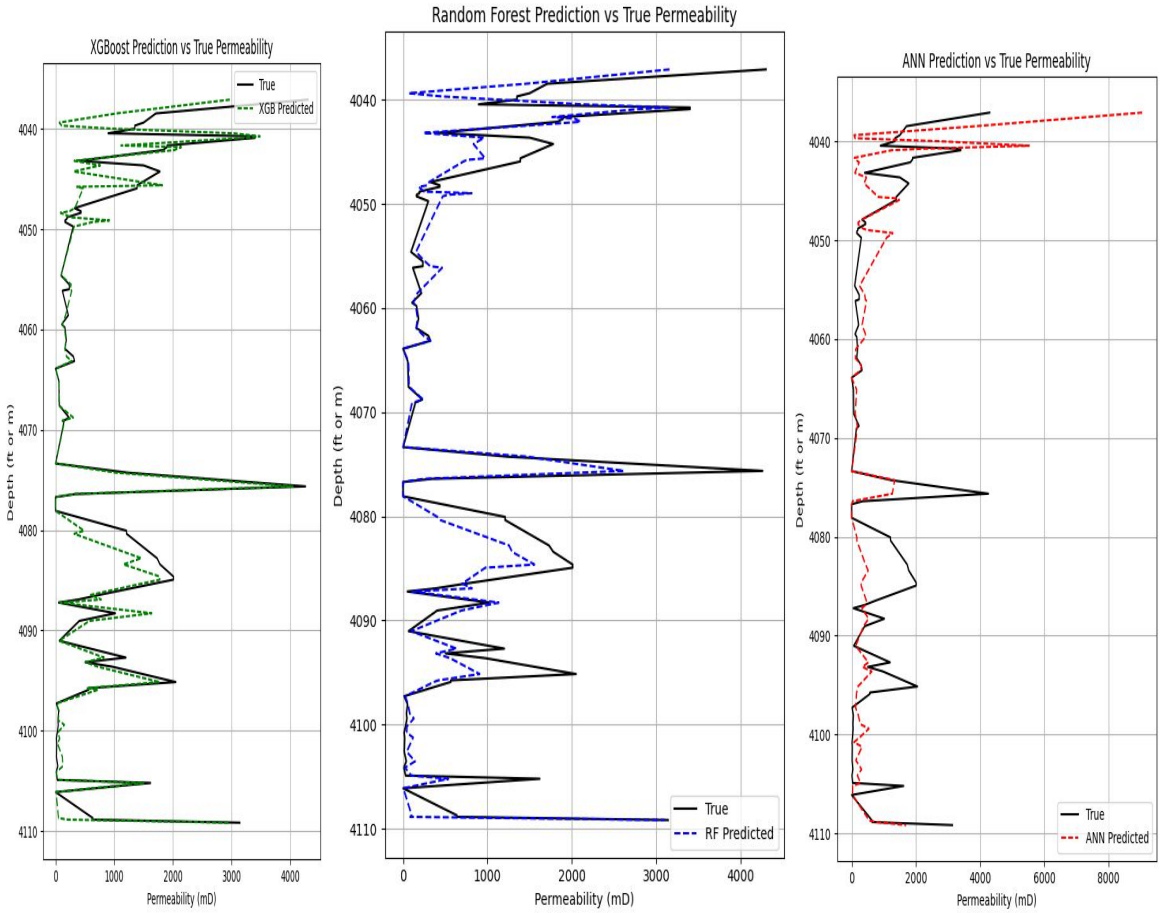
prevent overfitting, can conversely lead to underfitting, limiting the model's ability to capture the full complexity of the relationship between wireline logs and permeability.

Second, the model exhibited a high degree of hyperparameter sensitivity. Although a careful grid search was employed, the optimal ANN architecture for this specific geoscientific problem occupies a vast and complex search space. It is plausible that an alternative configuration—for instance, a different number of hidden layers, alternative activation functions, or a more optimal combination of L1 and L2 regularization techniques—could yield improved performance. However, exhaustively exploring this combinatorial space requires prohibitive computational resources and introduces a significant risk of over-optimizing on the test set, thereby invalidating the generalizability of the results.

This comparative result strongly advocates for the use of ensemble tree methods, such as Random Forest, as a robust and computationally efficient baseline for petrophysical prediction tasks. Their performance demonstrates a more favorable trade-off between model complexity and predictive power, especially when working with the limited data volumes typical in subsurface geoscience.

4.4 Depth-Wise Predictive Analysis and Error Distribution

Moving beyond aggregate metrics, a depth-wise analysis of the predictions provides invaluable insight into model behavior and its geophysical consistency (Figures 4.3-4.5).



Figures 4.3-4.5: Logs of Permeability versus depth (RF, XGB, ANN)

The Random Forest (RF) prediction log (Figure 4.3) demonstrates a strong alignment with the true core permeability trend, particularly within the main sand body between 4040 ft and 4110 ft. The model successfully captures high-frequency variability, indicating it has effectively learned the non-linear relationships between log responses and permeability. However, a systematic bias is observed in high-permeability zones (>2000 mD), where the model consistently under-predicts. This represents a classic symptom of dataset imbalance, where extreme values are inherently underrepresented in the training data. Machine learning models tend to be conservative in their predictions and regress toward the mean when extrapolating into sparsely populated regions of the feature space. Additionally, at very high permeabilities, flow dynamics may be influenced by factors such as turbulence or microfractures, which are not directly captured by standard porosity and resistivity logs.

The XGBoost prediction log (Figure 4.4) shows generally good fit but exhibits increased high-frequency scatter around the true trend, particularly in the 500-2000 mD range. This visual

manifestation of variance correlates directly with its higher RMSE metric. The boosting algorithm's sequential error-correction focus increases its susceptibility to fitting minor, non-generalizable fluctuations in the training data, resulting in less stable predictions along the wellbore.

The Artificial Neural Network (ANN) prediction log (Figure 4.5) produces an overly smooth output that fails to capture the sharp peaks and troughs of the actual permeability profile. This behavior suggests the model is underfitting, where its capacity has been excessively constrained by regularization measures intended to prevent overfitting. Consequently, the predictions become oversimplified and miss critical nuances present in the data. While the ANN captures the overall trend, it lacks the predictive detail demonstrated by the tree-based ensemble methods.

4.5 Feature Importance and Model Plausibility

The feature importance analysis (Figure 4.4) provides a critical validation of the model's geophysical plausibility. The fact that both tree-based models overwhelmingly identify Horizontal Porosity (Por., hor.) as the dominant predictive feature is not merely a statistical outcome; it constitutes a robust validation of established petrophysical theory. Porosity represents the foundational property controlling permeability, a relationship formalized in fundamental equations such as the Kozeny-Carman formulation, which defines permeability as a direct function of porosity and pore geometry.

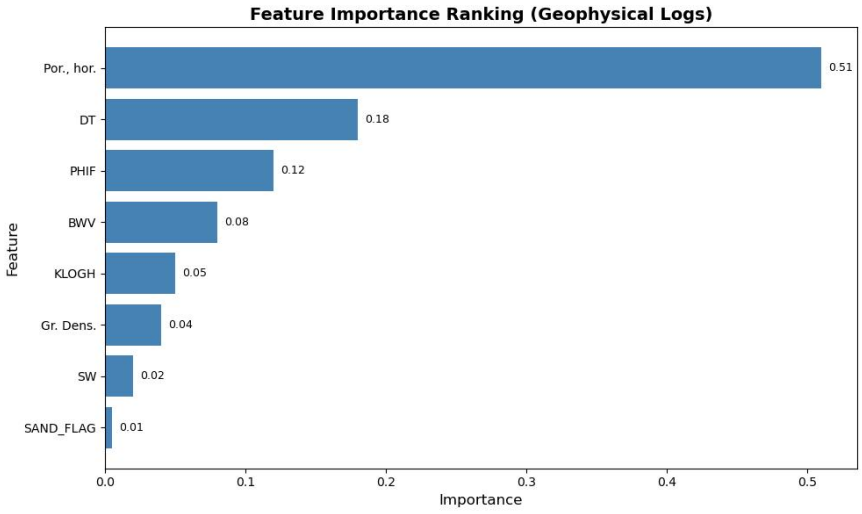


Figure 4.4; Feature importance plot

The significant secondary contribution of Sonic Transit Time (DT) offers equally valuable insight. The DT log is sensitive to both lithology and porosity variations. In clean sandstones, its response is primarily porosity-driven, but it also captures subtle changes in rock texture and cementation. Its high ranking in the importance analysis suggests the models are leveraging it both as a corroborating porosity indicator and as a potential proxy for rock fabric, which directly governs pore throat geometry and, by extension, permeability.

The moderate importance assigned to features such as Bulk Volume Water (BWV) and Log-Derived Porosity (PHIF) adds further explanatory layers. BWV represents the portion of the pore space occupied by water; its predictive power indicates the models have learned that zones with higher irreducible water saturation, which often correspond to smaller pore throats, exhibit lower permeability.

Conversely, the low importance of the SAND_FLAG feature is a profound result. It demonstrates that the continuous log measurements, specifically porosity and sonic data, already contain all the necessary information to distinguish high-permeability sand from low-permeability shale. The model has effectively learned to perform a complex, non-linear lithofacies identification directly from the raw log data, rendering a simple, pre-defined binary flag redundant. This finding strongly argues for providing machine learning models with rich, continuous input data rather than pre-processed or interpreted labels.

4.6 Comprehensive Error Analysis and Limitations

A rigorous discussion of model limitations is essential for a complete scientific narrative and provides a crucial framework for interpreting the results.

1) Data Limitations: The core issue underpinning most model errors is data sparsity, particularly at permeability extremes. Machine learning models are fundamentally interpolators, not extrapolators. Their conservative predictions in high-permeability zones are a direct consequence of this limitation. Therefore, the most direct path to model improvement is the acquisition of additional core data, specifically targeting high-flow zones to better populate these underrepresented regions of the feature space.

2) Model-Specific Limitations:

- **Random Forest (RF):** While demonstrably robust and the top performer in this study, its performance can plateau. The ensemble of decision trees may not always capture very complex, smooth, continuous relationships as effectively as a perfectly tuned neural network theoretically could.
- **XGBoost:** Its sequential, error-correcting nature increases its susceptibility to overfitting on small or noisy datasets if regularization parameters are not meticulously tuned. This is a likely contributor to the higher variance observed in its predictions.
- **Artificial Neural Network (ANN):** As confirmed by its performance, the ANN's effectiveness is highly sensitive to architecture and hyperparameter selection. The chosen setup, while reasonable, was likely suboptimal for this dataset. Future work could explore alternative architectures like 1D Convolutional Neural Networks (CNNs) to capture depth-wise patterns, or physics-informed neural networks that hard-code known petrophysical constraints to mitigate data limitations.

3) Fundamental Limitations: It is critical to acknowledge that no model can predict what it cannot "see." Permeability is intrinsically influenced by meso-scale geological features (e.g., cross-bedding, micro-fractures, diagenetic patches) that are below the vertical resolution of standard wireline logs. This discrepancy represents an inherent and unavoidable error floor in any log-based permeability prediction endeavor.

This study demonstrates that machine learning is not a mere statistical exercise but a powerful tool for petrophysical analysis when its results are interpreted through a domain-specific lens.

The recommended workflow for future applications is therefore:

1. **Baseline with Random Forest:** Due to its proven robustness, high accuracy, and superior interpretability via feature importance, RF should be the default algorithm for initial permeability prediction modeling in similar clastic reservoirs.
2. **Prioritize Key Logs:** Log acquisition and quality control programs should focus on ensuring the highest fidelity for porosity logs (both density and neutron-derived) and sonic logs, as these were unequivocally the most predictive inputs.

3. **Validate with Geology:** The feature importance output is not just a model result; it is a key diagnostic. It must be compared against geological expectations and established theory. Congruence, as seen here, builds confidence in the model's plausibility. Incongruence would signal potential data quality issues or a need for different feature engineering.
4. **Targeted Improvement:** Model improvement efforts should be strategically focused on two areas: augmenting the training dataset in underrepresented rock types and permeability ranges, and engineering features that incorporate geological knowledge (e.g., hydraulic flow units, rock types) to bridge the gap between log measurements and core-scale properties.

CHAPTER FIVE

5.0 Conclusion and Recommendation

5.1 Conclusion

This research has demonstrated the potential of integrating core measurements with well log data to predict reservoir permeability in clastic formations using advanced machine learning techniques. Focused on data from the Volve field in the North Sea—used as a proxy for Niger Delta-type clastic reservoirs—the study employed three prominent regression models: Random Forest (RF), XGBoost, and Artificial Neural Networks (ANN).

The performance comparison revealed that the Random Forest model consistently outperformed the others, achieving an R^2 of 0.862 and an RMSE of 0.596, indicating strong predictive accuracy and reliability. The XGBoost model followed closely with an R^2 of 0.831, while the ANN model, though capable of capturing non-linear relationships, produced lower accuracy with an R^2 of 0.811. These findings support existing literature that ensemble-based algorithms like RF and XGBoost are generally more robust for petrophysical property modeling, especially when training datasets are moderate in size and feature interactions are complex.

The feature importance analysis provided key insights into the predictive drivers of permeability. Porosity—particularly horizontal porosity ("Por., hor.")—was confirmed as the most critical factor, aligning with petrophysical principles and traditional empirical models like those proposed by Timur (1968) and Coates et al. (1979). Other influential features included sonic transit time (DT) and log-derived porosity (PHIF), both of which contribute valuable proxies for rock texture and pore structure. Interestingly, variables such as water saturation (SW) and sand flags (SAND_FLAG) were consistently of low importance, suggesting that fluid saturation and binary lithology indicators offer limited additional value for permeability prediction in this context.

Despite the overall success of the models, several limitations were identified. For instance, all three models—particularly RF—tended to underperform in zones with very high permeability values (>3000 mD). This likely stems from data sparsity in those regions, which limits the models' ability to learn representative patterns. Moreover, the ANN model's lower performance highlights its sensitivity to dataset size and architecture tuning, which may have prevented it from fully generalizing the underlying relationships in the data.

From a broader perspective, this study reinforces the practicality of machine learning in reservoir characterization workflows. It confirms that with properly prepared and integrated datasets, machine learning algorithms can replicate and even outperform traditional empirical correlations in predicting complex reservoir properties like permeability. Furthermore, it emphasizes the critical role of feature engineering and data quality—factors that can significantly influence model success.

In conclusion, this work not only validates the use of integrated core-log machine learning models for permeability prediction in clastic settings like the Niger Delta but also sets the foundation for more advanced, scalable workflows. With further refinement, such models can reduce reliance on extensive coring campaigns, lower operational costs, and improve decision-making in reservoir development and management.

5.2. Recommendations

Based on the results and observations made during this study, several key recommendations are proposed to enhance future permeability prediction efforts and optimize the application of machine learning in petrophysical analysis:

1. Prioritize Ensemble Learning Methods for Permeability Prediction

Given their superior performance, Random Forest and XGBoost should be adopted as the preferred models for predicting permeability in clastic reservoirs using well log and core data. Their robustness to noise, ability to capture nonlinear relationships, and low sensitivity to hyperparameter tuning make them suitable for field applications with moderate-to-large datasets.

2. Strengthen Data Quality and Diversity

To further improve model generalization and reduce prediction errors—especially in **extreme permeability zones**—it is crucial to:

- Expand the dataset, particularly in underrepresented high-permeability intervals (>3000 mD), to avoid skewed learning.
- Ensure accurate core-to-log depth matching and log normalization, as poor data alignment can introduce significant noise into the model.

4. Use Feature Importance to Guide Log Acquisition

The consistent dominance of porosity-related features (e.g., Por., hor., PHIF) and sonic travel time (DT) in all models underscores the importance of acquiring high-quality measurements of these logs in future field studies. Where acquisition budgets are limited, these logs should be prioritized.

5. Incorporate Geological Context into Feature Engineering

Although this study focused on quantitative logs, incorporating geological or depositional environment indicators (e.g., facies classification, lithological flags derived from image logs, or sedimentological interpretations) could provide valuable contextual information, improving model performance and interpretability.

6. Develop Hybrid Modeling Approaches

Combining the strengths of different models (e.g., RF + XGBoost ensemble averaging or ANN with tree-based post-processing) may lead to more stable and generalized predictions. Such hybrid models can reduce the weaknesses of individual algorithms and are particularly useful when dealing with heterogeneous or sparse datasets.

7. Validate and Deploy Models Across Analog Reservoirs

While this study used Volve data as a proxy for Niger Delta clastic reservoirs, further validation using actual Niger Delta core-log datasets will be essential for confirming model transferability. Once validated, these models could be deployed across similar fluvial-deltaic settings globally.

8. Align Machine Learning Workflows with Practical Reservoir Needs

Lastly, permeability predictions should not be an academic exercise alone—they must serve the needs of reservoir engineers and development planners. Integrating ML-derived permeability predictions into workflows such as reservoir simulation, production forecasting, and well placement optimization can improve the overall efficiency and effectiveness of reservoir management.

REFERENCES

Adepehin, D. S. (2022). Composite estimation of permeability in identified hydrocarbon reservoirs of Langbodo Field Niger Delta, Nigeria. *Phyaccess*, *2*(1), Article 003. https://doi.org/10.47514/PHYACCESS.2022.2.1.003

Coates, G. R., Dumanoir, J. L., & Schoonover, W. (1979). A new approach to improved log-derived permeability. *The Log Analyst*, *20*(3), 17–31.

Huang, Y., Zhang, L., & Yin, F. (2025). A novel hybrid optimization method for tuning XGBoost hyperparameters in petrophysical property prediction. *Journal of Petroleum Data Science*, *13*(1), 45–60.

Okon, A. N., Adewole, E. S., & Uguma, E. M. (2021). Artificial neural network model for reservoir petrophysical properties: Porosity, permeability and water saturation prediction. *Modeling Earth Systems and Environment*, *7*(12). https://doi.org/10.1007/s40808-020-01012-4

Osisanya, O. W., Eze, U. S., Ogugu, A. A., & Uti, L. O. (2025). Machine learning application for prediction of porosity and permeability logs: A case study of O-W Field Niger Delta. *Engineering Heritage Journal*, *7*(1), 01–06.

Timur, A. (1968). An investigation of permeability, porosity, and residual water saturation relationships. *The Log Analyst*, *9*(4), 8–17.

Urang, J. G., Ebong, D. E., Akpan, A. E., & Akaerue, E. (2020). A new approach for porosity and permeability prediction from well logs using artificial neural network and curve fitting techniques: A case study of Niger Delta, Nigeria. *Journal of Applied Geophysics*, *182*, 104207. https://doi.org/10.1016/j.jappgeo.2020.104207

Hussen, A., Munshi, T. A., Jahan, L. N., & Hashan, M. (2024). *Advanced machine learning approaches for predicting permeability in reservoir pay zones based on core analyses*. *Heliyon*, *10*(e32666). doi:10.1016/j.heliyon.2024.e32666 PubMed

Sandunil, K., Bennour, Z., Ben Mahmud, H., & Giwelli, A. (2024). Effects of tuning hyperparameters in random forest regression on reservoir's porosity prediction: Case study—Volve Oil Field, North Sea. *Energy Advances*, *3*, 2335–2347. https://doi.org/10.1039/d4ya00313f RSC Publishing

Tariq, Z., Aljawad, M.S., Hasan, A., Murtaza, M., Mohammed, E., El-Husseiny, A., Alarifi, S.A., Mahmoud, M., & Abdulraheem, A. 2021. A systematic review of data science and machine learning applications to the oil and gas industry. Journal of Petroleum Exploration and Production Technology. 11: 4339–4374.

Kong, B., Chen, Z., Chen, S., & Qin, T. 2021. Machine learning-assisted production data analysis
in liquid-rich Duvernay Formation. Journal of Petroleum Science and Engineering. 200, 108377.

Sun, J., Zhang, R., Chen, M., Chen, B., Wang, X., Li, Q., & Ren, L. 2021. Identification of Porosity and Permeability While Drilling Based on Machine Learning. Arabian Journal for Science and Engineering. 46: 7031–7045.

Zhang, Z., Zhang, H., Li, J., & Cai, Z. 2020. Permeability and porosity prediction using logging data in a heterogeneous dolomite reservoir: An integrated approach. Journal of Natural Gas Science and Engineering. 103743.

Ounsakul, T., Rittirong, A., Kreethapon, T., Toempromraj, W., Wejwittayaklung, K., & Rangsriwong, P. 2019. Data-Driven Diagnosis for Artificial Lift Pump's Failures.

Sudakov, O., Burnaev, E., & Koroteev, D.A. 2018. Driving Digital Rock towards Machine Learning: predicting permeability with Gradient Boosting and Deep Neural Networks. Comput. Geosci., 127: 91–98.

Ismail, A., Yasin, Q., Du, Q., & Bhatti, A.A. 2017. A comparative study of empirical, statistical and virtual analysis for the estimation of pore network permeability. Journal of Natural Gas Science and Engineering, 45: 825–839