Integrated Approach on AI/ML assisted Log reconstruction methodology for estimation of the missing/non-recorded logs

A Thesis submitted to the

UPES

For the Award of

Doctor of Philosophy

in

Petroleum Engineering

By

Bishnu Kumar

May 2025

Supervisor

Dr. Sunil Kumar Khare

Co-Supervisor

Dr. Tanupriya Choudhary

Department of Petroleum Engineering and Earth Sciences

Energy Cluster, School of Advanced Engineering

UPES

Dehradun – 248007: Uttarakhand, INDIA

Integrated Approach on AI/ML assisted Log reconstruction methodology for estimation of the missing/non-recorded logs

A Thesis submitted to the

UPES

For the Award of

Doctor of Philosophy

in

Petroleum Engineering

By

Bishnu Kumar
(SAP ID: 500087073)

May 2025

Supervisor

Dr. Sunil Kumar Khare
Professor
Department of Petroleum Engineering and Earth Sciences
UPES

Co-Supervisor

Dr. Tanupriya Choudhary

Professor
School of Computer Science
UPES

Department of Petroleum Engineering and Earth Sciences

Energy Cluster, School of Advanced Engineering

UPES

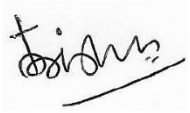
Dehradun – 248007: Uttarakhand, INDIA

DECLARATION

This is to certify that the thesis titled “INTEGRATED APPROACH ON AI/ML ASSISTED LOG RECONSTRUCTION METHODOLOGY FOR ESTIMATION OF THE MISSING/NON-RECORDED LOGS” has been authored by me. It presents the research conducted by me under the supervision of Prof. Sunil Kumar Khare and Prof. Tanupriya Choudhury.

To the best of my knowledge, it is an original work, both in terms of research content and narrative, and has not been submitted elsewhere, in part or in full, for a degree. Further due credit has been attributed to the relevant state-of-the-art collaborations with appropriate citations and acknowledgments, in line with established norms and practices.

Date: May’ 2025

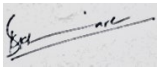


Bishnu Kumar
SAP ID#500087073
Department of Petroleum Engineering and Earth Sciences
UPES, Dehradun

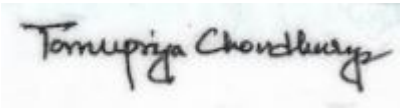
CERTIFICATE

It is certified that the work contained in the thesis entitled “INTEGRATED APPROACH ON AI/ML ASSISTED LOG RECONSTRUCTION METHODOLOGY FOR ESTIMATION OF THE MISSING/NON-RECORDED LOGS” by Bishnu Kumar SAP ID#500087073 has been carried out under our supervision and that has not been submitted elsewhere for a degree.

Date: May’ 2025



Dr. Sunil Kumar Khare
Professor
UPES
Dehradun



Dr. Tanupriya Choudhury
Senior Associate Professor
UPES
Dehradun

ABSTRACT

Name of the student: BISHNU KUMAR SAP ID: 500087073
Degree for which submitted: DOCTOR OF PHILOSOPHY (PhD)
Department: PETROLEUM ENGINEERING AND EARTH SCIENCES
Thesis title: Integrated Approach on AI/ML Assisted Log Reconstruction Methodology for Estimation of the Missing/Non-Recorded Log
Thesis supervisors: Prof. DR. SUNIL KUMAR KHARE & DR. TANUPRIYA CHOUDHURY
Month and year of thesis submission: May' 2025

One of the important procedures in the hydrocarbon exploration and production process is petroleum well logging, which achieves in-depth information pertaining to subsurface geological formations in drilling. The technique describes how different tools and sensors are used to collect data relating to the properties of the rock and fluids that are considered essential in the characterization of petroleum reservoirs. Wells' logs give data about the important parameters that include porosity, permeability, fluid saturation, and lithology. This information will then be used by the geoscientist in constructing accurate subsurface models that are very crucial in the estimation of hydrocarbon reserves, optimization of drilling strategies, and analysis of viability for different extraction methods. From the analysis of the data acquired, petroleum engineers can make good decisions on reservoir management, leading to improved operational efficiency. A major drawback of petroleum well logging is that the process often suffers from a variety of technical problems that can seriously degrade data quality and redundancy. Consequently, poor quality data acquisition may result from equipment failure. Other adverse borehole conditions may also influence logging operation: unstable formations, for example, can cause significant disturbance to logging activities; similarly, environments characterised by elevated pressure conditions can affect logging operations. Such factors greatly enhance the uncertainties in the reconstructed datasets from incomplete or noisy data sets during reservoir analysis.

Such challenges, therefore, affect the geological model developed by geoscientists, which may in turn affect the managerial decision related to resource extraction and the overall reservoir management. To enhance the outcomes of well logging in the oil industry, new innovative solutions such as advanced data processing and some machine learning techniques are required to adequately address these issues. The proposed framework is based on the principle of data

liberation, where raw data of well logs is fetched from various disparate sources. Volve field dataset has been chosen from the Norwegian sector of the North Sea. Log data used in the current work included compressional slowness, density, gamma ray, neutron porosity, and resistivity logs. A robust and accurate AI/ML workflow requires a consistent and standardized set of well logs, hence a lot of time was spent in correcting the inconsistent and improper data formats including empty entries. The AI/ML framework designed in course of this work has rebuilt the reconstruction process, thereby systematically solving the complex problems.

The first major step in the designed framework, as part of current work, was the data liberation process. In this step, collection and consolidation of well log data was done from various sources, which included databases, field records, and the surveys done by geologists. In particular, the Volve Field dataset was a good case study on account of excellence of completeness and richness of data. However, even collected data from Volve Field, showed some inconsistencies resulting from the differences in recording time, standards, formats and different gathering methods. Hence, subsequent stages in the reconstruction process designed in this thesis, required consistency and standardization in the dataset. Techniques for this data set's standardization is being presented in this thesis to make it ready for processing and analysis by learning algorithms. This research study used an integrated framework that applied sophisticated machine learning methods in reconstructing incomplete well logs to further enhance subsurface exploration and reservoir characterization. This framework has several core components, including the following: data liberation, preprocessing, outlier detection, feature selection, clustering and machine learning based prediction. Each of these components was designed to address a particular challenge that is encountered within traditional reconstruction of well log methods, always ensuring that there is a systematic yet efficient approach to the handling of data. This iterative process of the framework becomes of utmost importance so that one can ensure the accuracy and reliability of the reconstructed logs, thus playing a decisive role in making critical decisions for hydrocarbon exploration.

Data liberation is the initial step in the methodology. Here, raw, plain well log data is freed from all source locations. In this step, data is standardized to ensure consistency and compatibility of machine learning algorithms since the dataset poses a challenge in the form of varying formats and missing entries. Once data has been freed, the next phase is the comprehensive preprocessing of data through cleaning, transformation, and preparation, which tends to enhance the quality and reliability of data. Thereby, removal of irrelevant or incorrect entries compromising the model prediction is done using various outlined techniques. Removal

of constant continuous values is done by data clipping. For outlier detection, three distinct machine learning algorithms were deployed: Local Outlier Factor (LOF), One-Class Support Vector Machine (SVM), and Isolation Forest (IF). All of them have strengths in outlier detection and comparing them for performance helped determine the best way to handle the dataset for the well logs. Outlier handling ensures that the dataset is accurate and representative of the modeled geological conditions.

After preprocessing, feature selection was achieved using various methods; for instance, to determine appropriate attributes to reconstruct the results. Several techniques, such as the correlation coefficient, Principal Component Analysis, and variance threshold plots, were applied for analysis and reduction of the dimensionality of the dataset. These kept maintaining computation efficiency and improved model performance by retaining only essential features. Finally, with feature selection, initiation started for the clustering process, by which the wells were grouped together by using Gaussian Mixture Model (GMM), where they exhibit similar characteristics geologically. This was important because it allowed the reconstruction method to take advantage of regional geological patterns, thus improving the machine learning model. The final steps were the computation of machine learning prediction and interactive visualization. It utilized Random Forest in reconstructing data for missing well log information since it is robust in handling complex, high-dimensional datasets. The methodology used covers the specifics of the Random Forest model, including decision tree construction, with a point of emphasis on hyperparameter tuning and cross-validation to ensure optimal model performance. Above that, uncertainty estimation and validation ensured reliability in predictions made, which required the use of interactive visualization tools as an aid to stakeholders for interaction and validation of reconstructed logs. These tools allow users to navigate and visually analyse the reconstructed logs, thereby making the collaboration between geoscientists and data scientists possible and ensured that the final outputs are mathematically valid and geologically useful for practical applications in subsurface exploration.

The performance of the machine learning model was tested with multiple performance metrics and validation strategies. Cross-validation was applied to the model for missing data performance, ensuring that it is generalized well beyond the dataset used for training. The following work discusses the importance of hyperparameter tuning, particularly when optimizing parameters of Random Forest optimization: Number of Trees and Maximum Depth. Grid search and random search were covered as ways to identify the best combinations of model's parameters to acquire optimal performance for prediction. I have discussed

performance metrics at length, such as mean absolute error and root mean square error, to convey how well these metrics represent the model accuracy in terms of the reconstruction of well logs. It is critical that visualization and validation are a fundamental component of the framework so that the reconstructed logs are credible and useful for applications. Details about interactive visualization tools has been given based on how such tools empower users to inspect and evaluate the reconstructed logs effectively. This contribution especially underlines the need for visual user interfaces that enable seamless exploration of results to be made available to foster communication and collaboration between domain experts and data scientists. Since the framework is grounded on a collaborative basis, interdisciplinary contact, log reconstruction must be both mathematically correct and geologically relevant for subsequent analysis and decision making.

To demonstrate the approach, present work consists of a full case study using the Volve Field data set. The entire reconstruction process-from data liberation to the application of machine learning models-is guided with a case study to illustrate how the framework can effectively reconstruct the missing well logs. The accuracy and reliability of the AI/ML-assisted approach is analysed by using reconstructed logs and actual geological data for comparison. More importantly, the benefits of well log reconstruction to reservoir management is explained in detail, including the improvement of data quality in reservoir characterization, optimization of hydrocarbon extraction strategy and operational risk minimization. The discussions based on the work and its suggestion during the work in future could also explore the wider impact of AI/ML-assisted well log reconstruction on subsurface exploration with the potential scope for other applications in carbon capture and storage (CCS) and geothermal energy exploration. This work aimed at illustrating the capabilities of AI and ML in resolving geoscientific problems, thereby demonstrating their potential to revolutionize routine practices in subsurface analysis. Future research direction would involve three subjects: more complex reconstruction tasks through deep learning models, integration of real-time well log data for continuous monitoring, and expanding the framework to include larger and diversified datasets. Ultimately, in conclusion, it is emphasized that a transformative potential of AI/ML exists in improving the efficiency, accuracy, and reliability of well log reconstruction for further informed decisions in subsurface exploration and management.

ACKNOWLEGDEMENT

I would like to express my deepest gratitude to my revered PhD supervisors Dr. Sunil Kumar Khare and Dr. Tanupriya Choudhury for their constant guidance and encouragement throughout my research journey. I would like to specially acknowledge Dr. Sunil Kumar Khare's support provided irrespective of challenging time zone differences (India vs Europe Vs North America) along with his expertise, encouragement, availability has profoundly shaped this thesis. I would also like to acknowledge support provided by Dr. J. Devaraju & Dr. Arvind Sharma, mentors in shaping up this work. Unwavering guidance, support, encouragement along with insightful feedback provided by my supervisors & mentors has been indispensable throughout my doctoral pursuit.

I am deeply thankful to UPES for furnishing the resources and facilities necessary for conducting this research. The administrative support & guidance provided by Dr. Pankaj Kumar, Dr Nilanjana Banerjee, Dr Adesh Kumar, Mr Yogesh Kumar Sharma and Mrs Shipra Tiwari during various stages of this journey has been integral to the culmination of my PhD journey.

I would also like to acknowledge my sincere gratitude and thanks to M/S Equinor, Norway for providing access to complete datasets of Volve Oilfield, a part of North Sea for research, study and development purposes. This is a unique initiative by Equinor and has immensely helped me completing the current work.

Additionally, I am grateful for the support extended by SLB, which enabled me to pursue my doctoral studies. I wish to express my appreciation to Ms. Patricia Cejas Manceron, Global AI & INNOVATION FACTORI Manager for her constant support & encouragement. I would also like to acknowledge support provided by Mr. Ivan Velikanov, Global Operations Support Manager and Mr. Sushant Shekhar, Domain Data Scientist, whose technical insights were invaluable to this work.

Saving the best for last, I am forever indebted to the most loving and supportive spouse, children and my parents who are responsible for making me a decent being. They have seen me burn the midnight oil and stood by my supposedly whimsy decisions. I love you Bela, Adya, Avighna, Babu jee and Maa I dedicate all my life's work to you.

Dedicated to my family & parents

CONTENTS

DECLARATION iii
CERTIFICATE iv
ABSTRACT v
ACKNOWLEGDEMENT ix
CONTENTS xi
LIST OF FIGURES xv
LIST OF TABLES xvii
LIST OF ABBREVIATIONS xviii
LIST OF SYMBOLS xix
CHAPTER 1 1
INTRODUCTION 1
1.1 BACKGROUND 1
1.2 Understanding Oilfields: A Critical Component of the Energy Sector 1
1.4 Interconnection Between Oilfields and Wireline Logging in Resource Development 6
1.5 BACKGROUND OF WIRELINE LOGGING 7
1.5.1 PROCESS OF WIRELINE LOGGING 8
1.5.2 PREPARATION FOR WIRELINE LOGGING 9
1.5.3 Deployment of Wireline Tools 10
1.5.4 Data Acquisition and Measurement 11
1.5.5 Interpretation of Logged data 13
1.5.6 Post-Logging Operations 13
1.6 Wireline Logging Services 14
1.7 Wireline Logging Challenges 14
1.8 Challenges Experienced in Logs Derived from Wireline Logging 16
1.9 Missing Logs as the Problem of Wireline Logging 16
1.10 Problem encountered in Log Reconstruction 17
1.11 Importance of AI/ML for Operational Log Data in Critical Systems 18
1.12 Problem Statement 19
1.13 Motivation / Need for Research 20
1.14 Research Objectives 20

1.15 Research Question 21
1.16 Research Methodology 21
1.18 Theoretical framework 23
1.19 Organization of Thesis 26
1.20 Summary 27
CHAPTER 2 28
Literature Review 28
2.1 Background of the study 28
2.2 Overview of Wireline Logging 28
2.3 Difficulties in Petrophysical log data 29
2.3.1 Challenges 29
2.3.2 Research questions (RQ) 31
2.3.3 Data and Methodology 31
2.3.4 Search Strategy & Methodology 32
2.3.5 Inclusion and exclusion procedure for article selection 33
2.3.6 Results & Discussion 34
2.4 Log Reconstruction Techniques 44
2.5 Well log Reconstruction using ML 45
2.7 Well Log Reconstruction Using Deep Neural Networks 55
2.8 Well Log Reconstruction Using AI Techniques 59
2.9 Critical Observation 64
2.10 Gap of the research 65
2.11 Industrial Tools 66
2.12 SWOT Analysis 67
2.13 Summary 68
Chapter 3 69
Research Methodology 69
3.1 Reconstruction of Well Logs Using the PROPOSED INTEGRATED Framework of Machine Learning Techniques 69
3.1.1 Key Logs for Machine Learning Analysis 71
3.2 Dataset Description 75
3.2.1 Attributes of Dataset 76
3.2.2 Data Volve Field Dataset- Well LOG RECONSTRUCTION 78
3.2.3 Dataset Overview and Availability 78
3.2.4 Wellbore Platform 80

3.2.4.1 AI Platform 80
3.3 Data Pre-processing 81
3.3.1 Data Cleaning and Transformation 82
3.3.2 Removal of Constant Continuous Values 84
3.3.3 Data Clipping 85
3.4 Detecting Outlier using THREE ML ALGORITHMS 86
3.4.1 Local Outlier Factor 88
3.4.2 One-Class Support Vector Machine 90
3.4.3 Isolation Forest 94
3.4.4 Comparison of Outlier Detection Methods for Well Log Data 97
3.5 Feature selection Using various Methods 99
3.5.1 Correlation Coefficient 100
3.5.2 Principal Component Analysis 104
3.5.3 Variance Threshold Plot 107
3.5.4 Feature Selection Using Correlation Coefficient, PCA, and Variance Threshold Plot 109
3.6 Clustering Process Post-Feature Selection 111
3.6.1 Clustering using Gaussian Mixture Models 111
3.7 ML Prediction Using Random Forest Algorithm 114
3.7.1 Random Forest Structure 115
3.7.2 Building Decision Tree 115
3.7.3 Hyperparameter Tuning and Cross-Validation in Random Forest 117
3.7.3.1 Cross Validation Procedure 117
3.7.3.2 Training and Data Preparation 118
3.7.4 Uncertainty Estimation and Validation 119
3.8 Interactive Visualization 120
CHAPTER- 4 122
Result and Discussion 122
4.1 Background 122
4.2 Analysis of Data Pre- processing 124
4.2.1 Evaluation of Cleansing and Transformation 124
4.3 Detection of Outliers - Analysis 128
4.3.1 Identifying LOG Outliers using LOF Method 128
4.3.2 Detecting LOG Outliers using svm Method 131
4.3.3 Detecting LOG Outliers using IF Method 133

4.3.4 Logs Outlier Detection USING THREE ALGORITHMS....136
4.4 Analysis of Wells and Variables Selection....139
4.5 FEATURES SELECTION ANALYSIS....142
4.5.1 Selection of Features using Correlation Coefficient....153
4.5.2 Selection of Features using PCA....155
4.5.3 Selection of Features using Lasso and Variance threshold....159
4.6 Analysis of Clustering using Gaussian Mixture Models....161
4.7 Evaluation of ML Training and Prediction....167
4.8 Performance Metrices....171
4.9 Comparison of Performance metrices....174
4.10 Discussion....177
Chapter- 5....180
Conclusion AND Future Works....180
5.1 Conclusion....180
5.2 FUTURE WORKS....183
References....186
List of publications....207
APPENDIX- 2....208
PLAGIARISM REPORT....208

LIST OF FIGURES

Figure 1: Petrophysical Analysis of data from well log 5
Figure 2: Wireless Logging 7
Figure 3: Wireline Logging Process 9
Figure 4: Deployment of wireline tools in an oil or gas well setting..... 10
Figure 5: General System Architecture and Workflow 22
Figure 6: Feed Forward Neural Network Architecture..... 24
Figure 7: Basemap of Volve Field. 25
Figure 8: 3D Map of Wells from Volve Field with reservoir Zone..... 26
Figure 9: 2D Base Map highlighted faulted base reservoir 26
Figure 10: Flow chart for Systematic Literature Review..... 32
Figure 11: Yearly distribution of published articles 33
Figure 12: Co-occurrence of various index keywords (Machine Learning)..... 35
Figure 13: Co-occurrence of various index keywords (Well Logging)..... 35
Figure 14: Co-occurrence of various index keywords (Artificial Intelligence)..... 36
Figure 15: Co-occurrence of various index keywords (Well Logs) 36
Figure 16: Co-occurrence of various index keywords (Published Year)..... 37
Figure 17: Co-occurrence of various index keywords (total co-occurrence) 37
Figure 18: Proposed Framework Block Diagram 70
Figure 19: AI/ML Based LogQC, editing and reconstruction workflow..... 72
Figure 20: Data outline- 4TB of data, Well Logs, Seismic, Production..... 79
Figure 21: Underlying AI platform workflow 80
Figure 22: Predicting outliers using LOF Algorithm..... 89
Figure 23: Detecting Outliers using One Class SVM..... 92
Figure 24: Outlier Detection Using Isolation Forest..... 96
Figure 25: Feature selection using Correlation Coefficients 102
Figure 26: Feature selection using PCA 106
Figure 27: Feature Selection using Proposed Methodology..... 110
Figure 28: AI/ML Based Data QC, Cleaning and Transformation workflow. 126
Figure 29: Cross plot of Density Vs Sonic Curve, showing Outliers identified in RED color using LOF method, marked in circles 129

Figure 30: Cross plot of Density Vs Sonic Curve, showing Outliers identified in RED color using SVM method, marked in circles 132
Figure 31: Cross plot of Density Vs Sonic Curve, showing Outliers identified in RED color using IF method, marked in circles 134
Figure 32: Outlier detection workflow with different methods. 136
Figure 33: ML QC and Well Selection workflow. 139
Figure 34: Feature Selection workflow comparing results from various methods highlighting the key results after the Feature Selection workflow was run to compare results from all 5 algorithms 143
Figure 35: Feature Selection process results after successful execution and Output was compared in the form of relevant_var. 146
Figure 36: Cross Plot of Sonic Curve (AC) along y-axis Vs Various input curves (x-axis) along with predicted trend line 150
Figure 37: Results from Corr-Coeff, for Feature Selection 153
Figure 38: Results from PCA for Feature Selection 156
Figure 39: Results from Lasso and Variance threshold for Feature Selection 159
Figure 40: Clustering workflow showing first pass results as displayed in cross-plot. 161
Figure 41: AC-DEN-NEU Cross plots highlighting identified clusters following GMM Clustering 165
Figure 42: Machine Learning Training and prediction workflow 168
Figure 43: ML Log Prediction results display in AC-NEU-DEN Cross plots 169
Figure 44: Performance Comparison of MAE with Existing Frameworks 175
Figure 45: Performance Comparison of RMSE with Existing Frameworks 176
Figure 46: Performance Comparison of R-squared with Existing Frameworks 177

LIST OF TABLES

Table 1: Key Characteristics and Operational Factors of Oilfields 2
Table 2: Data acquisition phase during wireline logging 11
Table 3: Challenges in Wireline Logging 15
Table 4: Issues in Log Reconstruction 18
Table 5: A list of the drawbacks of AI and ML algorithms 30
Table 6: Various combinations 33
Table 7: Summary of Co-occurrence analysis via VOSViewer 38
Table 8: Summary of the AI & ML methods via VOSviewer network analysis along with advantages/limitation to be used with well logs 40
Table 9: Comparative Analysis of One-Class SVM, Isolation Forest, and Local Outlier Factor (LOF) for Outlier Detection in Well Log Data 98
Table 10: Comparison of Total Vs Outlier datapoints through algorithms 128
Table 11: Performance Comparison with existing methods 174

LIST OF ABBREVIATIONS

Abbreviation		Full Form
AI	-	Artificial Intelligence
AIC	-	Akaike Information Criterion
BIC	-	Bayesian Information Criterion
E-Step	-	Expectation Step (in EM Algorithm)
GMM	-	Gaussian Mixture Model
k-NN	-	k-Nearest Neighbors
LOF	-	Local Outlier Factor
LRD	-	Local Reachability Density
ML	-	Machine Learning
M-Step	-	Maximization Step (in EM Algorithm)
PC	-	Principal Component
PCA	-	Principal Component Analysis
PDF	-	Probability Density Function
RF	-	Random Forest
SVM	-	Support Vector Machine
XGBoost	-	Extreme Gradient Boosting
Z-Score	-	Standard Score or Z-Score

LIST OF SYMBOLS

Symbol		Definition
x	-	Original log value
$log(x)$	-	Logarithmic transformation of x
z	-	Z-score of the log value
μ	-	Mean of the log values
σ	-	Standard deviation of the log values
V	-	Variance of log parameter values
N	-	Number of observations
$\bar{x}$	-	Mean of the log parameter values
y	-	Transformed value (logarithmic)
b	-	Base of the logarithm (commonly 10 or e)
x'	-	Clipped log value
$Lower\ Limit$	-	Lower boundary for acceptable log measurements
$Upper\ Limit$	-	Upper boundary for acceptable log measurements
$LOF(X)$	-	Local Outlier Factor for point X
$N_{k(x)}$	-	Set of k-nearest neighbors of x
$lrd_k(x)$	-	Local reachability density of x
Ω	-	Weight vector defining the hyperplane
V	-	Parameter controlling outliers and margin of hyperplane
$\emptyset(x_i)$	-	Mapping of input data to higher-dimensional space through kernel function
ρ	-	Offset from the origin, controlling placement of hyperplane
$s(x,n)$	-	Score of point x in isolation tree
$E(h(x))$	-	Expected path length for point xxx in isolation tree
$c(n)$	-	Average path length for dataset of size n
r_{xy}	-	Correlation coefficient between x and y
PC_k	-	k-th principal component
w_{ki}	-	Weight for feature iii in component k
Z	-	Data matrix of principal components (new features)

X	-	Original data matrix
W	-	Matrix of eigenvectors (principal components)
λ_i	-	Eigenvalue corresponding to i-th principal component
$P(x)$	-	Probability density function for point xxx
π_k	-	Mixing coefficient for cluster k
μ_k	-	Mean of Gaussian distribution for cluster k
$\sum k$	-	Covariance matrix for cluster k
γ_{ik}	-	Probability of data point belonging to cluster k
L	-	Likelihood of the model
AIC	-	Akaike Information Criterion
BIC	-	Bayesian Information Criterion
$f_n(x)$	-	Output of n-th decision tree for input xxx
$\hat{y}$	-	Final predicted output
$X_{cluster}$	-	Features of a particular cluster
$y_{cluster}$	-	Target for a particular cluster
$\widehat{y_1}\ and\ \widehat{y_2}$	-	Predictions from the top two clusters
$\omega_1\ and\ \omega_2$	-	Normalized probabilities of the sample belonging to the two clusters

Integrated approach on AI/ML assisted Log reconstruction methodology for estimation of the missing/non-recorded logs

CHAPTER 1
INTRODUCTION

1.1 BACKGROUND

Originally the oil and gas industry require data captured with adequate and consistent information to meet the safety and hydrocarbon productivity. The basis of this data collection framework is logging that offers valuable information on one or the other aspect of reservoir conditions, drilling activities, and equipment. In characterizing the geological structure especially, the formation, in assessing the progress of the drilling operation and the overall well integrity, logs are crucial factors that determine the success of exploration and production. Nonetheless, original and valuable log data sources contain some significant drawbacks: inaccurate, containing gaps or containing partial information because of technical faults, failures of equipment, or delays in transmission (Gupta & Shah, 2022). These deficits raise concern about data persistence, which is desired in real-time, operationally oriented decision-making environments. Decisions taken with such incomplete datasets (gaps not filled up, data missing for some intervals) can lead to inaccurate and sub-optimal drilling locations thereby risks rise in operational costs. However, given these challenges, the continuous developments in the Artificial Intelligence (AI) and/or Machine Learning (ML) technologies offer new and viable approaches towards reconstructing missing log data. To elaborate, AI and ML when used can help review existing sets of data and perform improvements, prediction, and missing data fills as a result of efficient patterns recognition of log data. Not only does it increase the quality of the data collected but also provides decision-makers with better information to assist them to make right choices in terms of drilling that serves best interest of enhancing some measures of safety and, at the same time ensuring sustainable operations in such a keener market. The combination of these technologies outlines a critical stage in how the oil and gas sector addresses the role of data, moving toward a more robust and adaptive organizational model (Rui, 2017).

1.2 UNDERSTANDING OILFIELDS: A CRITICAL COMPONENT OF THE ENERGY SECTOR

Oilfields are distinct regions on the earth's surface where petroleum and natural gas resources exist beneath the surface of the earth where organic material undergoes heat and pressure

change over millions of years in the production of hydrocarbons. Oilfield can be classified as large, deep and heterogeneous, and on this basis, it affects the kind of approach and tools used in its exploration and production. Exploration and development of oilfields have a central place in the world, providing hydrocarbons that are used as fuel and energy and are a basis for a number of chemical products that are essential for today's intensive industrial processes (Ashry, 2022).

This work aims to understand factors that govern the operational effectiveness and the feasibility of the oilfield development in relation to the reservoir type, drilling strategies, and the available extraction techniques. Modern methods and tools as 3D seismic survey and EOR methods allow to extract hydrocarbons with lowest possible negative impact on the environment (Giem, Sheiretov, & Couble, 2018) (Al-Jaberi & Al-Mayyahi, 2018) (Carpenter, 2023). In addition, production of hydrocarbons from various oilfields is checked and evaluated constantly to monitor the health of the reservoir and the efficiency gain during business, allowing prudent management to act appropriately. In this context, petrophysical data grows into a critical issue, as it characterises the physical and chemical aspects of the reservoir rocks and fluids. Amidst the growing levels of sophistication of oilfield endeavours, incorporation of higher levels of data acquisition and data analysis tools, especially in the field of artificial intelligence and machine learning is critical in optimizing the decision-making processes, increase recovery factors and the suitable economic viability of continued oilfield operations in the contemporary and competitive energy industry.

Table 1: Key Characteristics and Operational Factors of Oilfields

Parameter	**Description**	**Example Value**
Oilfield Name	Name of the oilfield	Example Oilfield A
Location	Geographic coordinates	Latitude: 34.1234° N
		Longitude: 118.5678° W
Reservoir Type	Type of geological formation	Sandstone
Size of Oilfield	Area covered by the oilfield (in km^2)	150 km^2

Depth of Reservoir	Average depth of the oil reservoir (in meters)	2,500 meters
Hydrocarbon Type	Type of hydrocarbons present	Oil and Natural Gas
Initial Production Rate	Initial production capacity (in barrels per day)	5,000 barrels/day
Current Production Rate	Current production capacity (in barrels per day)	4,200 barrels/day
Cumulative Production	Total hydrocarbons extracted (in barrels)	1,200,000 barrels
Estimated Reserves	Estimated remaining recoverable reserves (in barrels)	10 million barrels
Production Techniques	Techniques employed	Conventional, Enhanced Oil Recovery (EOR)
3D Seismic Imaging Utilization	Use of 3D imaging technology for exploration	Yes
Average Reservoir Pressure	Average pressure in the reservoir (in psi)	3,000 psi
Average Temperature	Average temperature of the reservoir (in °C)	80 °C
Monitoring Frequency	Frequency of reservoir monitoring	Monthly
Petrophysical Data Collected	Types of petrophysical data collected	Porosity, Permeability, Saturation
Environmental Impact Assessment	Overview of environmental assessments	Regularly conducted

Technological Integration	Integration of AI/ML in operational processes	Yes
Current Challenges	Key challenges faced in oilfield management	Equipment maintenance, Data integrity
Future Development Plans	Planned projects for improvement	New well drilling, Technology upgrades

The characteristics of the sample survey and table 1 identify those published metrics or operational factors that are of value for quantifying basic attributes of oilfields as well as practicing certain peripheral elements or approaches to oilfield development and operation. The table presents what is regarded from geographic analysis as the field's name and location or area which designates the position of the oilfield (Nguyen, Gosine, & Warrian, 2020). It describes the reservoir type to define the structure that hosts the hydrocarbons, including how and how much the company will extract it. The size of the oilfield measured in square kilometres gives an indication of the magnitude of the operation and the depth of the reservoir characterizes the average vertical separation to the subsurface of interests which in turn affects the drilling methods and technologies (Sales, 2019) (Tjioe, et al., 2022) (Wheater, Huyton, Miller, Gerrity, & Ford, 2018). In addition, the hydrocarbon type appears next; it separates it between oil and gas, crucial in different markets and extraction techniques. Primary measures are reported including the measure of the initial production rate and current production rate which gives an insight into the efficiency of the productivity of the oilfield over same period of time and its effectiveness in bringing up the trapped hydrocarbons. The cumulative production precisely illustrates the overall quantity of hydrocarbon that has been produced within the period of existence of such a field while estimated reserves are an important source of information regarding how viable the field could be in the future. Production techniques are emphasized where standard drilling or boosted oil recovery is applied. Literally, the 3D seismic imaging utilization displayed highlights exploration advancement's significance that the oil industries apply in enhancing extraction solutions based on subsurface mapping (Cheng, et al., 2019).

Moreover, as highlighted in subsequent sections, there are other critical measures like average reservoir pressure & average temperature both affecting production potential & plans. Monitoring frequency as to how often the amount of health of the reservoir is checked is also

highlighted so that it may be understood that the checking process should go on constantly to ensure that the production is at optimal levels. The sort of petrophysical data acquired includes porosity, permeability and saturation among others, they play an important role in the determination of the physical attributes of the reservoir in relation to resource management decisions. They include the environmental impact assessment which highlights the industry's sensitivity to the environment and its effects. Lastly, technological integration, for which the operational AI and machine learning are included in the book notes the change of process to data-loaded decision making.

It is unanimous that petrophysical sub-surface data is the primary workhorse of the oilfield development as they provide the framework for defining reservoir properties and, thus, hydrocarbon recovery (Hanson, Hyson, Munro, & Wheater, 2024) (Manakhayev, et al., 2020) (Jerram, 2019). In the past, the oil and gas industry has often required acquisition of data that covers years or even decades to evaluate the state of reservoirs and the performance of petroleum systems. Such data collection is important in reaching conclusions on exploration of potential well sites, production plans, and overall resource development. However, the cost associated with gathering this data is fairly steep; researchers have said that it costs from US Dollars One Hundred Thousand to One Hundred and Fifty Thousand dollars to acquire just one log curve (Sircar, Yadav, Rayavarapu, Bist, & Oza, 2021). Few expenses can add up so fast especially when undertaking a project that involves several logging runs and extensive raw data collection.

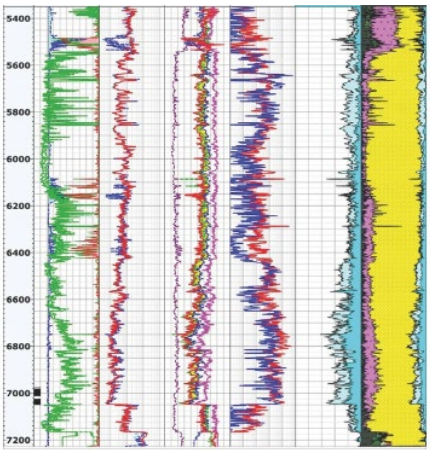


Figure 1: Petrophysical Analysis of data from well log

As over time oil economics changes and the fluctuating character of the oil and gas market has increased the awareness that data acquisition may not always be a given, instead it has to be earned. Therefore, it has evolved more towards gaining specific, high-quality data more than

starting extensive logging campaigns that go on for hours or days at a time (Li, Yu, Cao, Tian, & Cheng, 2021). This selective acquisition strategy will be an effort to maintain cost effectiveness while at the same time fulfilling the need to acquire the petrophysical data, which enables assessment of reservoir performance without severely impacting company's financial resources. While the process of selective data acquisition is difficult, there is a clear recognition that quality petrophysical data is essential in making key operational decisions and consequently to successful oilfield development. Therefore, solutions like state-of-the-art artificial intelligence and machine learning are gaining importance in addressing the need for the generality of data with the realities of the modern oilfield environment (Kulananpakdee, 2018) (Szabó, Kilik, & Dobróka, 2023).

1.4 INTERCONNECTION BETWEEN OILFIELDS AND WIRELINE LOGGING IN RESOURCE DEVELOPMENT

Wireline logging is an essential tool in the exploration and production activities of oilfields as it parameterizes relevant data required in the management of the fields. During wireline logging operation: specialized tools with multiple attached sondes are placed into wellbores through a wireline, this enables measurement of physical properties of underlying rocks and formations thereby helping engineers and petrophysics to interpret different geological and reservoir parameters for example, formation resistivity, porosity, and acoustic velocity. These parameters are crucial for evaluation of hydrocarbon presence and on the basis of the nature of the reservoir and the fluid in conjunction with the composition of the reservoir rock helps in well positioning to maximize the production and optimize drilling cost (Mohammadpoor & Torabi, 2020). The logging results impact the subsequent well completion approach as well as recovery technique and hence the total effectiveness of the oilfield. In addition, new technologies such as artificial intelligence and machine learning when applied in wireline logging procedures improves the manner and analysis of logs hence their ability to reconstruct lost data to enhance the accuracy of the reservoir descriptions. When oil fields are compounding and economic pressure for improved well productivity is escalating, wireline logging moves to the front and center as the support structure for ongoing determination and assessment for reservoir performance (Ngo, 2018) (Elnaggar, Kassab, & Abass, 2018) (Siregar, 2024) This relationship demonstrates the current and immediate need for fresh techniques in log data reconstruction, asserting that the proposed thesis on the use of AI/ML techniques in wireline logging is not only appropriate but necessary for the ongoing and future advancement of oilfield development and stewardship.

1.5 BACKGROUND OF WIRELINE LOGGING

Wireline logging is electrical logging which is a vital procedure used in oil and gas industry for determination of subsurface geological structures. They can be dated as far back as the beginning of the 20th century when the requirement for comprehensive subsurface data emerged during the exploration and spudding of oil wells (Alghamdi, Yateem, Rajeh, & Kadem, 2018). In the past, drilling had mostly employed surface data and geological maps in an effort to gain a clear picture about the structure of sub terranean revolutions, which were usually inadequate to explain the subsurface formations. As the companies extended more efforts towards concern more exploration, it was found that; there is a need of more accurate and timely way of getting the Wells information and the characteristics of the surrounding geology (Birt, Filiptsova, Ryan, & Hopper, 2019). The initial wireline logging tools were introduced in the 1920s with the primary method being based on electrical resistive measurements to establish rock characteristics. In recent years, the utilization of the technology has advanced considerably, and so have logging methods and sensors that can quantify many natural and chemical characteristics (Shuangfang, 2018). These enhanced subsurface images have aided in exploration and production trends and reservoir management in terms of improving our understanding of the nature of subsurface formations in the form of lithological logs, porosity, permeability, and fluids content profiles as shown in the figure 1.

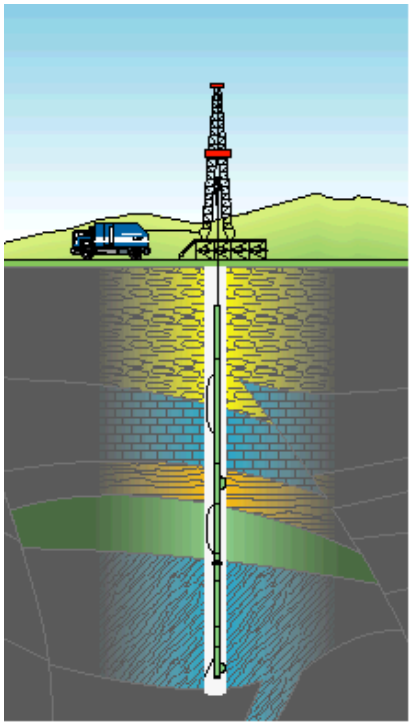


Figure 2: Wireless Logging

Wireline logging works by lowering down an array of complicated apparatus along the wellbore through a wireline cable which acts as a backbone, as well as a signal conductor. The information gathered through the logging operation is highly beneficial to geoscientists and engineers so that they build efficient well completions, extraction of hydrocarbons, and the management of reservoirs, among others. The increased ability to resolve data has improved geological models and performances thereby revolutionizing ways of predicting reservoir behavior and the driving of operational drilling. More recently, there has been a shift to the use of some of the newest technologies such as artificial intelligence (AI) and machine learning (ML) in the wireline logging procedures (Comyns, 2016). The former participations of these technologies facilitate the processing of huge data collected during logging undertakings, delivering improved understanding and prediction for planning events. However, wireline logging still has some problems such as problems associated with accuracy of data acquired, the effects that the logging tools have on the formation and the problem of real-time analysis. Solving these issues is crucial for the further development of the industry, stabilization of its already unstable sales, and the simplification of supply chains in an intensely competitive market (Shrivastava, Zhang, Yang, Forest, & Nir, 2021) (Chowdhury, Tanjil, Akter, Amin, & Pal, 2019) (Franco, Corona, AlMahroos, Kestring, & Oliveira, 2022).

1.5.1 PROCESS OF WIRELINE LOGGING

Wire-line logging is a critical method that is utilized in the OFS business to acquire comprehensive information about subsurface conditions throughout the different stages of the drilling and completion process of wells. The process starts with evaluation of the wellbore to which the drilling operations provide a path which is followed by wireline tools used (Franco, Corona, AlMahroos, & Kestring, 2022). After reaching the predetermined depth of a well, downhole tools with accurate protective casings which consist of thin strong cable for running logging tools required for calculating different geophysical parameters are inserted (Blinov, Filimonov, Ramatullaev, Tasemenov, & Saparova, 2020). These tools come fitted with devices that can capture features like the formation resistivity, porosity, density, and sonic velocity among others on a formation (Tawfeeq, Najmuldeen, & Ali, 2020). During the operation, the wireline is run up through the well while a number of logging tool's gather and transmit data to the surface and give engineers and geoscientists an insight of the well conditions as the operation proceeds (Alizadeh, Maroufi, & Heidarifard, 2018).

Figure 3: Wireline Logging Process

The data acquisition phase is especially important; the drilling crew as well as the logging engineers need to work closely to make certain the measurements are properly taken as shown in the figure 3. In most cases, different types of logging methods such as electrical resistivity logging, nuclear magnetic resonance logging, acoustic logging are used that reveals different features of the formation. The obtained logs afterwards are processed and analysed and raw data to readable formats using possibility of special programs and then the result is a visualization of logging (Siregar, Ramli, Turner, Prachayarittikool, & Iffwad, 2022) (Kelkouli, Arab, Amor, Kecili, & Mokhtari, 2019). These studies provide a broad understanding of reservoir quality and distribution of hydrocarbons zones, thus help engineers make sound decisions about well completions, and setting up production terms. In addition, wireline logging heavily contributes to augmenting the efficiency of drilling by proper management of the resources, eventually resulted in more safety aspect. Wireline logging is thus a critical tool that supports efficient understanding of subsurface formation and significantly enhances the prospects of exploration and production projects in the oil and gas business (Qiao, Shao, Bao, & Rong, 2017).

1.5.2 PREPARATION FOR WIRELINE LOGGING

A number of preliminary procedures have to precede the process of wireline logging: This includes deciding on the right logging instruments depending on the formation characteristics of the wellbore as well as the targets of the logging job. Engineers and geologists discuss what sort of data is necessary, some of which may be resistivity, porosity, density and sonic velocity (Mishra A. K., 2020). There is a variety of logging tools; electric logging tools, acoustic

logging tools, nuclear logging tools these tools are used to measure formation properties where; Next, the wellbore needs to be prepared for the logging Operation has to be carried out in a wellbore in order to obtain and record different parameters of the created well (Habeeb, 2023) (Abass, Teama, Kassab, & Elnaggar, 2019). It often covers the process of circulating or removing drilling mud, or other materials that may obstruct the tools which is in the well during logging. Where the borehole is very dirty, it may be necessary to employ other tools to have a clean well bore. It can sometimes be necessary to use conditioning fluids within a well in order to aid in logging especially in a wellbore environment that may be considered unfavourable or unstable (Lai, 2017).

1.5.3 DEPLOYMENT OF WIRELINE TOOLS

Once the well is prepared, there are several logging tools that are run down the well bore by use of a wireline cable. This cable, usually manufactured from high strength steel or fiberglass, plays the dual role of an electrical signal conductor from the surface to the downhole instruments and equipment as well as the device used to transmit such instruments and equipment into the well. It is, therefore, ensured that the deployment of the tools is controlled from the surface before being lowered to the required depths to take measurements (AlSunnary, AlMuslem, Al-Abdulmohsin, Bawazir, & Zaouali, 2019). As the logging tools descend, they are fitted with different types of sensors aimed at measuring the formation characteristics in real time. These sensors are designed to measure many specifications among them being electrical resistivity, gamma ray emissivity, sonic velocity among others (Al-Mudhafar, 2017). This information is transmitted back to the surface in real-time by the data acquisition system and engineers can track the progress of logging operation and take necessary corrections as shown in the figure 4.

Figure 4: Deployment of wireline tools in an oil or gas well setting

As we know, tools that are used wireline are essential in carrying out interventions as well as logging processes. It is a large cylindrical wireline tool employed for the activities including data acquisition and, principally, servicing. Thus, it is made up of several sections with different color, that is gold and green which represent a distinct functional area. They are connected to surface equipment's by a cable or wire hence can communicate. The wireline tool is provided with a cable which passes through pulleys/sheaves an arrangement made at the surface to control the movement of the tool. These are essential in data collection especially on subsurface formations, —instrument conveyance such as in unblocking, and in tool run without pulling the tubing. This arrangement also guarantees operation efficiency, and flows, making wells productive (Zakaria, Nawawi, & Salin, 2016).

1.5.4 DATA ACQUISITION AND MEASUREMENT

At this area, the logging tools enter the data acquisition phase of their operation. The tools collect several measurements as they move through the wellbore normally collecting information after a given time interval to give a clear image of the formations. These measurements are dependent on factors such as the kind of rock and quantity of fluids and other related environmental effects (Oliveira F. V., 2022) (Klitten & Agyekum, 2021) (Pootisirakorn, 2018). The information is used for the assessment of properties of the subsurface formations, definition of prospective for hydrocarbon accumulation layers and estimation of the well performance. During the data acquisition, it is possible that the logging tools use techniques that will increase measurement results accuracy. For example, in resistivity logging, the formation is energized by an electrical current which is used to determine the formation conductivity that may suggest the presence of hydrocarbons. Likewise, nuclear logging tools employed gamma ray to determine formation density and which is vital in determination of composition of the reservoir rock (Radwan, 2021).

Table 2: Data acquisition phase during wireline logging

Measurement Type	Description	Method	Purpose	Data Interval (ft)	Units
Resistivity	Measures formation's	Electric current introduced	Identifies hydrocarbon-bearing zones	0.5	Ohm-m

	electrical resistance				
Gamma Ray	Detects natural gamma radiation from formations	Gamma ray detection	Differentiates between shale and clean formations	1	API units
Density	Measures formation density	Nuclear (gamma radiation)	Assesses rock composition and porosity	0.5	g/cm³
Neutron Porosity	Measures hydrogen content	Neutron interaction	Estimates porosity and fluid type	0.5	Porosity (%)
Sonic	Measures sound wave travel time through formations	Acoustic wave propagation	Estimates formation porosity and lithology	0.5	µs/ft
Caliper	Measures borehole diameter	Mechanical arms	Detects borehole shape and wellbore integrity	0.25	Inches

Some of the measurements as shown in the table 2, which make up the wireline log data acquisition section, are explained in the table below and all are for the purpose of analysing the subsurface formations. While resistivity measurements pass an electric current through a formation in terms of conductivity, it assists in interpreting hydrocarbon-bearing zones. Gamma ray logging measures natural Gamma radiation which comes from the formation needed to differentiate between shale and improved, rock formations. Or data gathering on

subsurface formations, maintenance like clearing blockages, and tool deployment without removing the tubing. This setup ensures efficient, controlled operations, enhancing well productivity (Kausik, et al., 2016). Density logging gives details of the formation of the reservoir rock by employing nuclear methods to determine the density of the formation. Neutron porosity tools quantify for hydrogen in the rock and infer the porosity and the type of the invasion fluid, oil or water, among others. Sonic logging consists in generating sound waves that propagate through the formation and whose travel time is used to evaluate the formation porosity and lithology. Last but not the least, the caliper tool is used to determine the borehole diameter to check if the wellbore has been compromised at any section or not, and to maintain the integrity of data. Each is may be taken at different time to develop an integrated picture of both the wellbore and the formations within which it is located, thereby improving the appraisal and estimation of reservoir and well performance.

1.5.5 INTERPRETATION OF LOGGED DATA

Data interpretation is the second significant process in the acquisition of data after the data has been logged. This phase involves processing of the recorded measurements in order to come up with valuable information regarding the geologies involved. By using the log data, the geoscientists and engineers apply a number of software tools and modelling to create a detailed logs and cross-sectional model of the subsurface formation. In this case, interpretation may involve defining geological components such as the reservoir limits, and epic fractures and faults that support the hydrocarbon mobility (Garrouch & Al-Sultan, 2019) (Mishra, et al., 2019). Others include the advanced techniques like; Petrophysical analysis: tools needed for the quantitative evaluations of reservoir and formation that can contain reserves and specifies their characteristics like the porosity and permeability estimates which are useful in evaluating productivity in the well. The logs form a single part of an interrelated set of geological and engineering data that provides a clearer picture of the reservoir and helping when it comes to well completion and production planning (Tiab & Donaldson, 2024).

1.5.6 POST-LOGGING OPERATIONS

After the logging operation the tools are brought back from the wellbore and the results are checked to ensure all required data has been acquired and is correct. Sometimes, other logging may be necessary if there was a problem with the initial logging for instance if the trees got stuck during the logging process (Rossi, Martinez, & Useche, 2023). To achieve this, actual resource data have to be logged by consistent, accurate, documented, and validated measurements to reflect the subsurface conditions. This final logged data is then subjected to

validation and then incorporated into reports that will describe the construction of the geological formations. These reports are useful to engineers and in decision making after, in assessing future drilling activities, improving the production methods and in monitoring the performance of the reservoirs thereafter (Liu H. , 2017). Wireline logging therefore opens up other possibilities of helping with such decisions and indeed supports future carrying out of operations in the sustainable and efficient manner (Nordin, Azid, Ong, & Abdullah, 2024).

1.6 WIRELINE LOGGING SERVICES

Wireline logging also remains one of the most valuable methods used in the oil and gas industry, which fundamentally improves the acquisition of detailed information about the subsurface geology of areas adjacent to the wells. This technique entails use of specific logging tools that are placed in a wellbore using the wire line which enables engineers to obtain multiple important measurements with a high level of accuracy (Partouche, 2020) (Duthie, Saiood, Al-Anizi, & AlGhamdi, 2020). Among the above parameters, they include: formation resistivity to give information about the conductivity of the formation, porosity to reveal the void space within the rocks that holds hydrocarbons, sonic velocity to determine mechanical properties of formations and also are central to defining the production goals and managing the reservoir production in the long run. The detail that can be achieved with wireline logging provides by geologists and engineers a means of making comprehensive assessments of reservoir character, which is useful for more precise location of prospective areas for hydrocarbon accumulation and guidance for future drilling operations (Mo'men, Darwish, Abdelhady, & Essa, 2017). Also, wireline logging may be used in risk evaluation since detailed data that may help engineers avoid certain dangers during drilling are produced through this process. Wireline logging data used in the construction of geological models, effective ways of forecasting reservoir behaviour and the general economic returns realise in the exploration and production tasks.

1.7 WIRELINE LOGGING CHALLENGES

Nonetheless, wireline logging as an important tool in the oil and gas industry comes with its own challenges; these challenges are some that can severely affect the quality and credibility of the log data acquired. Among the most frequently reported problems is the signal loss which depends upon the electrical noise or improper cable connections (Fryklund & Stark, 2020). This kind of signal interruption can render some information untransformed completely missing within the log records and may hamper analysis big time. Tool failures are another

familiar issue; either the mechanical part of the logging tool fails or the particular component hampers the proper operation of a logging tool and lead to unlogged data or even wrong readings. Moreover, communication interfaced can be experienced when transferring data from the logging tools to the surface especially deep wells or when its environment is complex which makes the already problematic data continuity worse.

Table 3: Challenges in Wireline Logging

Challenge	Description	Impact
Signal Loss	Downhole equipment malfunctions	Missing depth intervals in the logs
Tool Failure	Logging tool stops functioning	Incomplete logs and the need for retesting
Transmission Delays	Delays in sending data from downhole to surface	Gaps in time-sensitive data logs
Borehole Conditions	Harsh environments affecting tool performance	Decreased accuracy of recorded data

In such conditions such as High pressure, high temperature as shown in the table 3, HPTH the probability of equipment failure predisposes logging tool because of the stresses experienced. Such conditions can result into long periods untouched, during which many logs go unrecorded leaving extensive gaps into the critical log information. However, the conditions in the borehole also affect the quality of data recorded during drilling processes. For instance, we may get contamination by mud filtrates that are used in the drilling process effect the measurement of parameters such as resistivity and porosity, among others. The leads to such degradation of data quality are not only detrimental to the logging operation at that point, but also hardly beneficial to the long-term management of the reservoir or for making informed decisions (Caporin & Fontini, 2017). Consequently, the management of these challenges is essential to the accuracy of data obtained from wireline logging and which in a way help to improve efficiency and safety relating to activities in the exploration and production of oil and natural gas.

1.8 CHALLENGES EXPERIENCED IN LOGS DERIVED FROM WIRELINE LOGGING

Problems associated with wireline logging might produce the following problems especially in terms of the formation evaluation accuracy and reliability being severely impacted by incomplete sections within the generated log data. In cases where there is limited log data or where there are gaps, the operator is confronted with inadequate information of the geological structures around the borehole. Such an absence of a wide range of integrated information hinders decision-making on key facets of the wells BHA and cluster design, hydraulic fracturing, and production control and monitoring (Barbosa & de, 2019). For example, if operators do not have all the information they need to determine the porosity and permeability of a formation they could make faulty decisions on where to drill a particular well or how to properly initiate the flow of the formation (Soeder, 2018). Moreover, constant and high-quality data is required for realization about hydrocarbon sustainability and for the formulation of strategies for extraction. An incomplete log data the operators undertaking his/her work of defining productive zones in the reservoirs and negates an operator's ability to accurately identify productive zones which may prove costly in terms of time and resources (Acemoglu, Lelarge, & Restrepo, 2020) (Agista, Guo, & Yu, 2018). In the long term, consequences of incomplete logs are followed to the capacity to model reservoirs behaviour in the proper way. This way, weak data accumulation can result in incorrect forecasts of recovery rates and in general – reservoir performance scales, through which operational costs skyrocket. These challenges, therefore, underscore the importance of finding improved approaches to rehabilitation and quality improvement of wireline logging data, to enable operators make accurate decisions for resource exploitation to reduce the costs of business risks.

1.9 MISSING LOGS AS THE PROBLEM OF WIRELINE LOGGING

An obvious severe drawback from wireline logging is the absence of log data in well evaluation and reservoir characterization, which causes a difficult approach to making proper decision in the oil and gas industry. Daily logs provide fundamental information of the subsurface that contains numerous parameters of rock, fluid and geological properties that define the prospect of hydrocarbon bearing beds. Operators are challenged by gaps of log data to interpret the characteristics of the reservoir and possibly misestimate resources and recovery plans. Simple extrapolation techniques that were formerly used to handle missing logs are generally ineffective because of the complexity and heterogeneity of most formations in sub-surface

geology. It should not be taken to mean that such straightforward techniques can sufficiently capture the real patterns in data and give direction to operators as they make their evaluation (Ahmed & Meehan, 2016).

However, the current practice of logging back to the well bore to retake measurements proves very difficult because of the following reasons; high expenditure, time consuming and the limitation of operations within drilling environment. Of course, such interventions entail massive costs and lead to prolongation of projects, which addresses the issue of operational efficiency. In order to meet this growing demand for ROC solutions, various AI and ML based advanced techniques have gained interest (Serra, 1983). These technologies proposed the capabilities of predicting the missing logs without physical interventions; in other words, machines can make better predictions by using the existing data for more comprehensive estimate. Through optimization employing modelling and prediction techniques AI/ML can improve the quality of characterization of the reservoir which will lead to better decision making for the efficient use of the resources within the reservoir as well as lower operational risks (Bell, 2018) (Blümich & Singh, 2018) (Civan, 2023). That shift to technical solutions also highlights the need for a solution to the problem of missing log data in wireline logging (Luthi, 2001).

1.10 PROBLEM ENCOUNTERED IN LOG RECONSTRUCTION

Although, research has significantly advanced in integrating AI and ML in log reconstruction the following challenges still emerge to hinder the overall efficiency of these advanced methodologies. There is first and prime problem of data imbalance in this case which means the logs generated from faulty situations are highly outnumbered by those coming from normal functioning of the system (Lai, 2022). This imbalance can cause the occurrence of biased models that do not capture correctly missing or anomalous data in their training to the extent of minimizing the efficiency of the predictions in real life incidences. The other difficulty relates to data noise – the presence of unrelated or incorrect information in the log entries. This noise can come from sensors and it has the effect of overlaying the fundamentals that the models use to make its predictions, which is why this noise must be removed. This type of noise therefore requires strong preprocessing methods to eliminate this kind of noise since they are detrimental to the feeding of clean data into AI/ML frameworks (Patterson, 2017).

Scalability is also another area of concern especially due to the increase in the log data generated in the oil and gas industry. AI/ML methods demand high computational capability

to process big data and hence have certain constraints when applied to situations where the analysis should be performed in a fast time. The requirement for potent processing framework, and efficient algorithms are also a constraint, typically in field operations where issues of resource constraint might arise (Ellis & Singer, 2008). The lack of interpretability of AI/ML models is the final challenge to be mentioned here. Most advanced deep learning paradigms are black box in nature, which mean that while engineers and decision makers receive a prediction or some model behavior, they cannot understand why that is the right or proper action to take. In other words, due to the lack of transparent data, it becomes rather challenging to incorporate AI/ML solutions in such high-stake environments as oil and gas operations. Therefore, to improve the utilization of log reconstruction approaches in subsurface exploration and production challenges need to be addressed and thus innovated for more suitable solutions as shown in the table 4.

Table 4: Issues in Log Reconstruction

Challenge	Description
Data Imbalance	Logs from normal operations dominate
Data Noise	Log entries contain irrelevant information
Scalability	High computational costs for large datasets
Model Interpretability	Difficulty understanding model predictions

1.11 IMPORTANCE OF AI/ML FOR OPERATIONAL LOG DATA IN CRITICAL SYSTEMS

Documents of logs form part of the base in numerous fields to track the performance of systems, forecast possible breakdowns, or as a tool in carrying out diagnostic examinations. This is especially the case in core industries like oil &gas, IT and health care this due to the fact that in these sectors, the risks of failure are much higher (Feldman, 2019) (Harter, 2019). These industries are such that the absence or limited amounts of log data may result in negative outcomes in many respects, including disruptions of processes, threats to safety, and lower efficiency of decision-making. For example, missing log data in the oil and gas industry becomes a disincentive to sound assessment of reservoir characterization hence reducing production effectiveness and propelling operational risks. In the same way, logs that are not

fully filled can affect patient monitoring of their treatment, as it can compromise individual patients' safety and healthcare (Ghosh, 2022).

Implementation of AI and ML in reconstructing logs is another compelling solution to the missing logs problems. These algorithms are generally in the forefront of analysing large datasets while identifying complex patterns; therefore they are capable of recreating missing log entries with considerable precision. This capability ensures the continuity of the data flow, allowing organisations to make informed and timely decision based on complete and accurate log information (Fan, Deng, Qiu, Zhao, & Wu, 2021). More importantly, it is possible to improve the reliability of the proactive monitoring tactics that help organizations prevent possible system failures. In operation-critical sectors where even minor operational decision can have drastic effect, adopting AI/ML assistive log reconstruction not only strengthens the operations and their safety but also forms the organizational safety culture. Hence by making sure that these decision makers have full and efficient logs, the organizations can move around the complex operational contexts with more certainty, thereby improving the results and the general sustained operational performance (Haustveit, 2020) (Jia, Tsau, & Barati, 2019).

1.12 PROBLEM STATEMENT

In the oil and gas industry, the missing petrophysical log data interpretation and processing have been handled historically using conventional approaches namely; K-Means Clustering, Angular-Based Outlier Detection, Supervised Neural Networks, and Self-Organizing Maps (SOM). These approaches, as used, have been adopted in common with several disadvantages. For example, many of them work as black boxes, that is, when a model makes a decision, it is often impossible to understand how it arrived at that decision. This absence of openness lends little credibility to the outcome, much less success in its usage in vital sectors like well log reconstruction (Aalsalem, Khan, Gharibi, Khan, & Arshad, 2018). Furthermore, these methods are anchored on simple correlations, which does not always reflect the contoured geological structures of subterranean formations. Therefore, they are typically large and complex and often need much human effort when it is being implemented. This results in low productivity, especially in large scale production, and adds the factor of low reliability. On the other hand, results attained by AI & ML technologies are instrumental especially in the supply chain, healthcare, and aviation industries particularly in automations, decision making, and improved operation. Nevertheless, the said applications have remained rather obscure in the O&G industry, despite the fact that the industry tends to require more sophisticated and automatised

solutions. Due to the specificity of subsurface data and potential influence of correct log interpretation on the further process of well drilling and reservoir management, the AI/ML application in this subject area is relevant (Nguyen, Gosine, & Warrian, 2020) (Sales, 2019). This thesis aims to overcome these limitations with an AI/ML-based approach for imputing missing log data; checking the log data for quality; and analysing logs for new nearby sections. The employment of the AI/ML algorithms will eliminate complicated and tiresome procedures and produce more precise and efficient results. This approach not only bypasses the drawbacks of the conventional approach but also has the potential to transform handling of petrophysical logs in the O&G industry in the same way AI has for other industries (Kato, 2020) (Kumar, Nagarajan, & Uchil, 2018).

1.13 MOTIVATION / NEED FOR RESEARCH

In subsurface exploration, wireline logging is a vital process in acquiring pertinent petrophysical data that facilitate an appreciation of the reservoir characteristics and subsequent proper determination of oil and gas to be exploited. However, logs collected through this method can have gaps or missing intervals due to practical difficulties or failures of the tools or due to circumstances prevailing within the well. Such gaps of data prove to be highly damaging in terms of reservoir modelling and decision-making processes that follow. Traditional techniques of log construction and QC include manual efforts which are tedious, time consuming and usually imprecise. Thus, there is a clear and increasing need for such processes to be automated using the technologies that the oil and gas industry is experiencing as a result of the Fourth Industrial Revolution. AI/ML prospects seem to better speed up, accuracy and reliability of log reconstruction potentially offering more precise predictions sections. This research fosters the urgent need for accurate, efficient, and automated approaches to reconstructing petrophysical logs and estimating their uncertainties by enhancing the subsurface reservoir management performance.

1.14 RESEARCH OBJECTIVES

- Evaluate the application of AI/ML techniques in log reconstruction using the Volve Field dataset, focusing on handling missing, noisy, and incomplete data in complex geological settings.
- Assess the benefits and limitations of key AI/ML methods such as LOF, One-Class SVM, Isolation Forest, GMM, and Random Forest for outlier detection, clustering, and well log reconstruction.

- ❖ Examine essential pre-processing and quality control steps, including data cleaning, outlier detection, and feature selection, to ensure high-quality input data for reliable AI/ML modelling.
- ❖ Develop strategies to minimize uncertainty and improve accuracy, leveraging advanced techniques like clustering-based prediction, cross-validation, and ensemble learning to enhance log reconstruction outcomes.

1.15 RESEARCH QUESTION

1) What is the current state of AI and machine learning (ML) techniques in log reconstruction?
2) What are the most commonly used AI/ML methods for log reconstruction, and what are their advantages and limitations?
3) What pre-processing and quality control steps are necessary to ensure accurate log reconstruction using AI/ML methods?
4) How can uncertainty be minimized and overall accuracy improved in AI/ML-based log reconstruction?

1.16 RESEARCH METHODOLOGY

The Integrated Approach on AI/ML Assisted Log Reconstruction Methodology for Estimation of Missing/Non-Recorded Logs research methodology is a multi-step methodology optimized for reconstruction with advanced data science techniques. Specifically, the process begins with data acquisition. Through the data from the well logs obtained from the Volve Field dataset, gamma ray, density, neutron porosity, and resistivity logs are conducted. These logs are crucial for subsurface characterization and lay a solid foundation for more advanced analysis. Once acquired, data liberation is conducted; that is, raw well log data extracted from proprietary formats or fragmented databases is transformed into standardized formats that can be fed into machine learning (ML) models. The next step is towards cleaning and transforming data where noisiness and outlier values together with missing values are handled using interpolation and normalization so that data becomes presentable for model development. Moreover, it will be uses to give back the corrected well logs together with outlier indicators on the original well logs and confidence intervals on predicted well logs as shown in the figure 5.

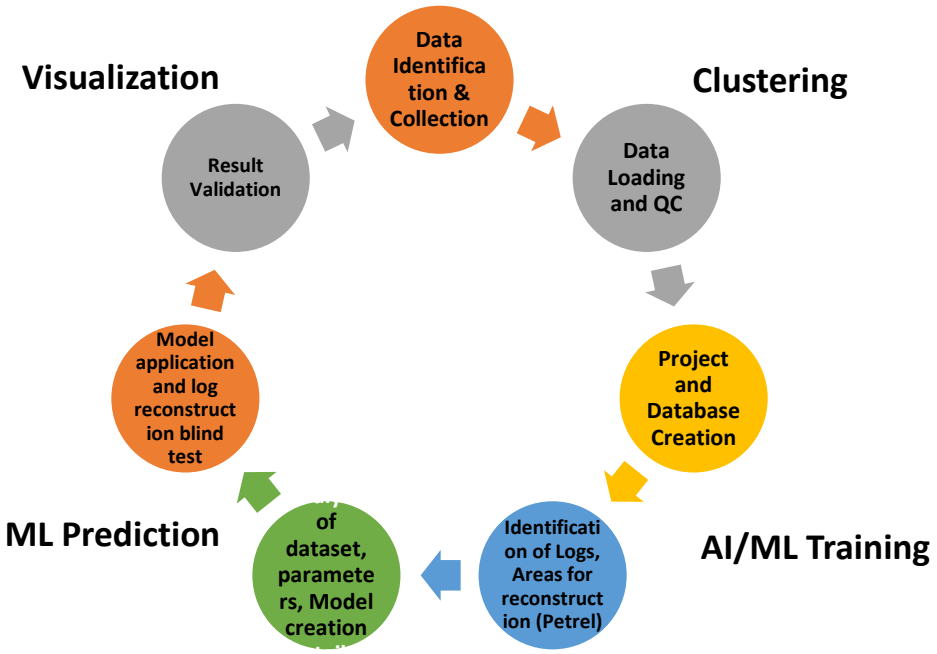


Figure 5: ***General System Architecture and Workflow***

The removal of outliers in the data uses the LOF algorithm, One-Class SVM, and Isolation Forest algorithms to eliminate skewness and inaccuracy in the model. The most predictive variables are selected based on their predictive power after the data cleaning step. Techniques like PCA or correlation analysis may be used in order to reduce the dimensionality or improve models as much as possible. Well clustering is performed next in which similar wells are grouped using cluster algorithms like Gaussian Mixture Models (GMM). Clustering allows the model to group wells with the same geological origin as one unit, enhancing the correctness of log reconstruction. Then, more complex machine learning algorithms, like Random Forests, which are optimal for heterogeneous data with missing values, were adopted. Using hyperparameter adjustment and cross-validation allowed us to optimize the predictive skills of the model, thereby reducing errors and making the model more robust. Lastly, the reconstructed logs are validated, and the authors construct interactive visualizations so that the reconstruction result can be verified by domain experts, decisions will be made, and further reporting of outcomes is facilitated in collaboration among data scientists and geologists for further investigation and refinement of a reconstruction model. The methodology, through this structured approach, aims at attaining a reliable and automated reconstruction of missing or non-recorded well logs, improving reservoir management and decision-making in complex geological settings.

1.18 THEORETICAL FRAMEWORK

Well log data has been the backbone in providing the reservoir properties that enables lithologies and fluids as well as oil and gas and water to be identified in the hydrocarbon industry. That is why there are many well logging methods; the physical properties of various types of rocks differ from each other. Well Logging techniques include formation evaluations techniques used in determining the lithology and reservoir characteristic of various formations, and production logging technique used in the control of well operation. At present, over thirty different methods of logging are possible. Other well logging technique include electrical and magnetic logging, radioactive logging, thermal well logging, acoustic and other logging method. Both are versatile and easy to store, run in the wellbore on wirelines. They quantify these parameters according to physical characteristics of rock e.g. electrical resistivity.

The application of Artificial Intelligence has resulted to increased innovation for the diverse fields and industries in the last years. Nevertheless, the implementation of the said technique in the Exploration and Production (E&P) Industry has not been frequently practiced. It was last decade that has experienced multiple cycle of Oil price swings up and down which has resulted Oil's cost efficiency optimization intendment being driven by AI/ML applications. Some of the differences between the above terms used frequently in the related domain are Artificial Intelligence, Machine Learning, Deep Learning and Neural Network have often been used interchangeably.

Neural network

Neural networks and ANNs in particular, are similar to the human brain by the computing algorithms used. On most basic levels, a neural network has four fundamental parts: it has inputs, weights, bias or threshold, and an output. Like linear regression, the algebraic formula would look something like this:

$$\sum_{i=1}^{m} w_i x_i + bias = w_1 x_1 + w_2 x_2 + w_3 x_3 + bias$$

Deep Learning

By deep learning, the term 'deep' refers to the number of layers present in a neural network. A neural network, if it has more than the basic architecture of three layers, wherein the inputs and

the output form part of it, then it would be perceived to run a deep learning algorithm. This is generally represented in the following figure 6.

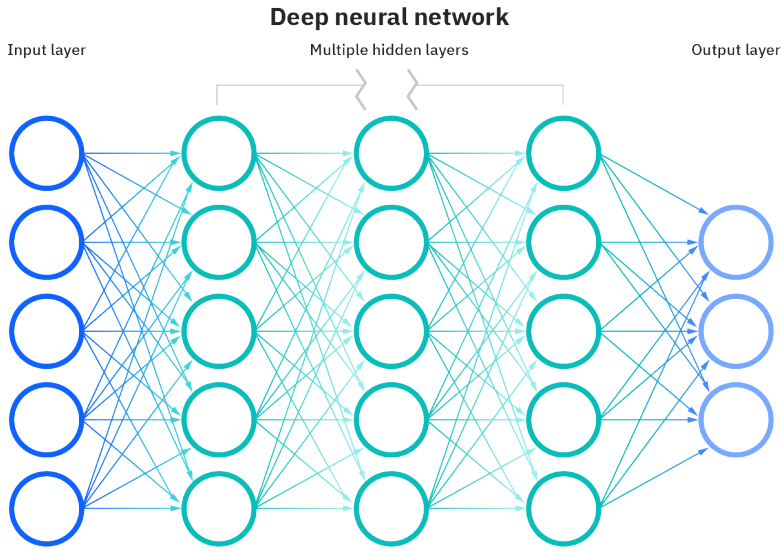


Figure 6: ***Feed Forward Neural Network Architecture***

Most deep neural networks are feed forward which implies that they process information from input to output with no often cycles or repetition as shown in the figure 6. But we can also train through back propagation that is start moving backward from output to input. Through back propagation one is able to compute the error linked to each neuron, and hence to correct it and fit the algorithm.

Machine Learning (ML)

Machine learning is a broader category of AI, in terms of approach and primary ways in which Machine Learning differs from Deep Learning is in the manner in which the system learns and the amount of data that algorithms of each category utilize. Deep learning on the other hand, speeds a lot of the feature extraction part and requires some of manual involvement. Large numbers are also allowed, and thus this machine learning is often referred to as 'scalable machine learning'.

Artificial intelligence (AI)

AI is used in this context to categorize equipment that imitates human intelligence. It can be applied to prospective, perform, and enhance area that has formerly done by humans, machine as well as process in E&P lifecycle such as repetitive log data acquisition. This work will discuss how artificial intelligence (AI) and machine learning (ML) have transformed the

process of handling missing log data in the oil and gas industry by integrating them into the process of log reconstruction. The efficiency of these AI/ML models is due to their flexibility that allows generalization from data and avoiding overfitting, a problem of more conventional methods. Thus, in addition to enhancing characteristics such as log data continuity and completeness, they also provide more accurate reservoir assessment and production plans. The application of AI/ML by the oil and gas industry would help the industry to move from the existing operational style that lacks comprehensive log data and loses significant amounts of money because of that shift. This shift also marks the need for technology in improving the reconstruction logistics or the process of getting even better techniques for the resource sector.

The data for this study has been collected from the Volve oil field located in the central part of the North Sea, at the southern part of the Norwegian sector as indicated in figure 7.

Figure 7: Basemap of Volve Field.

The field was first discovered in 1993, and the development plan was approved back in 2005. Then, field production came into full swing in early 2008 with a peak oil rate of 56,000 bbl/day. New wells had been drilled into the field until 2012–2013, and these new wells significantly added to a higher recovery rate and increased life of the field. The overall strategy for draining the structure was pressure maintenance with water injection, with the production wells placed high on the structure and water injectors at the flanks. The Volve field is described as a fault block structure with an initial estimation of 173 million bbl of oil in place. The reservoir is a small-dome-shaped structure and it is assumed to be generated due to the collapse of the salt

ridges adjacent to it during the Middle Jurassic age. Oil was produced from the sandstone of Middle Jurassic age in the Hugin.

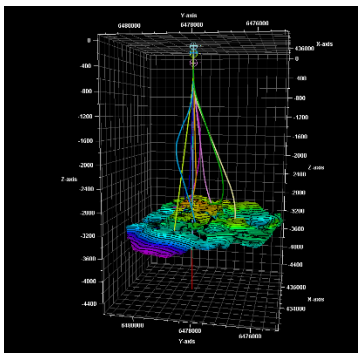


Figure 8: 3D Map of Wells from Volve Field with reservoir Zone

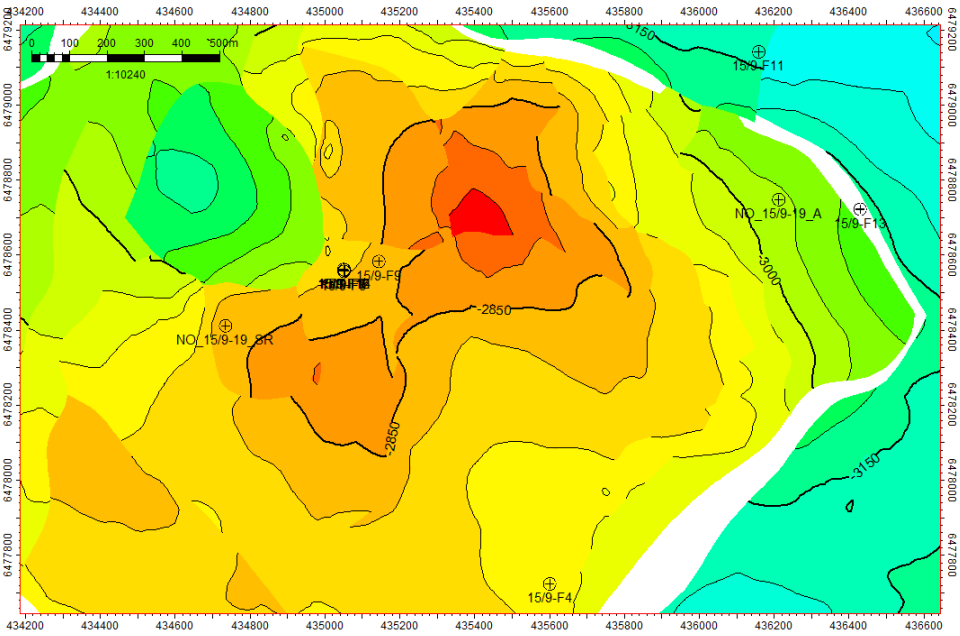


Figure 9: 2D Base map highlighted faulted base reservoir

1.19 ORGANIZATION OF THESIS

The thesis begins with Chapter 1: Background information, which outlines the overall subject of wireline logging, its relevance, use in the oil and gas industry and potential issues the occur when logs are missing or not recorded. This work describes why AI/ML will henceforth be utilized for log reconstruction, research objectives, and the structure of this thesis. Chapter 2: This Literature Review reviews academic articles concerning wireline logging, data difficulties, and the incorporation of AI/ML in log data, which are the gaps that your work fills.

Chapter 3: Section Methodology discloses the dataset that has been employed, data pre-processing, and AI/ML used to perform the reconstruction of missing log intervals. It also describes the process of model validation and the inclusion method of uncertainty estimation. Chapter 4: Results and Discussion highlights the outcomes of the applied AI/ML models and compares them to conventional approaches and highlights the importance of estimating uncertainty to logarithmic forecasting. Chapter 5: In Conclusion and Future Work section, overall findings of the research, its implication for the oil and gas industry, limitations of the research are presented, and future research recommendations are made. The thesis features the main text and more specifically the conclusion followed by the References section with all cited literature and the Appendices section for all extra material.

1.20 SUMMARY

This introduction focuses on the fact that petrophysical log analyses are at the heart of the oil and gas industry processes and performance success. Traditional methods to manage missing log data comprise K-Means Clustering, Supervised Neural Networks which are unsuitable for QI application due to their opacity and time-consuming nature. Identified key research areas are to automate QC; to develop ML models for log reconstruction; to predict logs for the adjacent sections and estimate uncertainty of the predictions. Data gathering, preparation, modeling, and quantifying model uncertainty are processes in the approach. This research aims at filling the gap of accurate and efficient methods in reconstruction of petrophysical logs to supplement existing methods in the reservoir management of the oil industry.

CHAPTER 2
LITERATURE REVIEW

2.1 BACKGROUND OF THE STUDY

AI and especially ML are the trends of recent years that have altered numerous industries providing unique solutions to numerous challenges. One of the applications of the technique is in the oil and gas industry and more to do with the analysis and interoperation of the petrophysical log data. Open and natural logs that contains very vital information about the subsurface formations such as porosity, permeability etc; are vital tools in reservoir characterizing, hydrocarbon defining, and for optimal production. Nonetheless, failure to record certain log intervals is extremely problematic and can cause further misconceptions that result in economic losses. Challenges associated with the application of conventional approaches to log reconstruction. AI and ML techniques provide an opportunity for an automating and improving the process of logs reconstruction. Since, the said technologies use algorithm which has the capability to learn from the previous patterns in data, the effectiveness of using such technologies to analyse the raw log data probably allows the researcher to identify the anomalies, inconsistencies and errors. Most importantly, deep learning models have demonstrated their capabilities of doing feature extraction and imputation of missing log sequences and predicting logs in the other regions where there are gaps. This research does not only enhance the existing method to determine the interpretation of the log but also provide real time decision in managing the reservoir (Li J. , 2018). Furthermore, the likelihood of log-predictions is improved by additional probabilistic models and ensemble learning techniques which allow uncertainty quantification in the course of data reconstruction. This is especially important in the oil and gas industry because unpredictability of subsurface conditions can greatly impact the plans and procedures in the industry. Consequently, this literature review seeks to investigate the state of the art of methodologies and their limitations before presenting AI/ML transformations in view of the quality control and reconstruction of petrophysical logs. This review will synthesize the existing literature to develop a systematic reference point from which to propose the novel AI/ML-supported log reconstruction methodologist, providing a starting point for future evolution of the essential field of research.

2.2 OVERVIEW OF WIRELINE LOGGING

One of the most influential procedures in the oil and gas industry is wire line logging which originates from the beginning of the century and has been developed every now and then

starting from the 1920s. Originally intended to collect geologic data about subsurface formations, wireline logging uses a variety of tools to gather information on physical properties of formations encountered while drilling. It involves running a series of sensors on a cable (or Wireline) to the wellbore while real-time data including resistivity, density, porosities and gamma-ray emissions can be obtained. Experience over the years has enable the use advanced logging tools to enable characterization of formation properties, and spectral gamma-ray logs to improve characterization of reservoir properties (Liu Y. , 2019) (Longde, 2021) (Masucci, Brusoni, & Cennamo, 2020) However, another key benefit is data acquisition whereby employment of the digital technology in integration enhances the acquiring of sharper and accurate data. Wireline logging has thus become very significant in the assessment of potential prospects for yielding hydrocarbons or other minerals, or decision making throughout the exploration of the reservoirs and production processes by the Petrophysicist and reservoir engineers.

2.3 DIFFICULTIES IN PETROPHYSICAL LOG DATA

Nevertheless, wireline logging, which is an essential component of reservoir description, has a number of limitations that can negatively affect petrophysical log data. One factor of concern is the presence of missing logs due to equipment breakdown, unfavourable situations or environment; or due to the intricate nature of the geology of the area which may hinder the acquisition of a complete record of subsurface conditions. Furthermore, noise, which results from interference from electrical signals, mechanical or fluid movement and can interfere with accurate assessment of the data collected. Discrepancy in data occurs due to differences in the logging instrument, manner of data collection, and even mistake which may occur during data analysis. Solving these problems necessitates reliable methods for data collection, which also involve methods of data integration and missing log computation (Nichol & Dhariwal, 2021). The combined method called Integrated Approach on AI/ML Assisted Log Reconstruction Methodology for Estimation of the Missing/Non-Recorded Logs that was developed in this thesis deals with these problems by utilizing superior artificial intelligence and machine learning methods, which lead to the improvement of the petrophysical analysis precision.

2.3.1 CHALLENGES

Log data is linked with a set of problems in a well which is usually bought by many service providers using multiple runs and passes in such well bore sections as are drilled using different mud parameters, bit diameter and hole size. Well log data may be missing or of poor quality for many reasons such as malfunctioning hardware, poor/washed-out hole circumstances,

equipment failure, loss of information due to inadequate data storage, and unfinished or incomplete logging resulting in the absence of log data for a region or possibly for an entire category of log.

For these reasons, a lot of exploration in recent years has focused on wire-line logging predictions. However, other have demonstrated certain disabilities and more possibilities. A major challenge to the application of machine learning in the energy industry is non generalizability of AI algorithms. Many designs struggle for differentiation if they failed to consider it at the time of their design as shown in the table 5.

Table 5: A list of the drawbacks of AI and ML algorithms

Limitation	**Reason**	**Solution**	**References**
Overfitting	A shortage of adequate information to be utilized in training.	Considering the proportion of the input points corresponding to all interconnections' net weights ([rho]).	(Andrea et al. 1991; Livingstone et al. 1997)
Coincidence	Obtaining an ideal match for a certain database by accident.	A discriminatory approach is used.	(Livingstone et al. 1993)
Overtraining	If the model's internal makeup is updated and the number of errors keeps dropping, the algorithm might get more complicated to match a specific data set.	The training approach known as "early ending" can be applied.	(Hossain 2018)
Data availability	Occasionally the information obtained is sparse.	Single-shot training is the process of enhancing a computerized model after it has been previously learned on a comparable collection.	(Weyrauch 2016)

Interpretability	The findings are affected by the cumulative effects of each of the model linkages rather than just the particular links in the scenarios.	Locally understandable system and unbiased justifications The technique of modified multiplicative designs.	(Shabbir et al. 2018; Bas et al. 2016)
Generalization	Degradation of the framework in conditions that are distinct from those employed to construct the initial model.	It is necessary to use greater assets to train fresh datasets.	(Virgina 2018; Ramamoorthy et al. 2018)
Bias	Black-box simulations are biased by their very characteristics.	Utilizing disturbances without affecting the framework.	(Samek et al. 2018)

2.3.2 RESEARCH QUESTIONS (RQ)

This comprehensive investigation of the literature attempts to find various released and pertinent publications about petrophysical drilling log quality control, editing, and reconstructing for absent or unrecorded sections using AI and ML. The aforementioned statements on issues are the main focus of this research's major goal:

RQ1. What is the current status of application of AI and ML based techniques for Log reconstruction?

RQ2. What are the applicable methods along with their benefits and limitations?

RQ3. What are various considerations like pre-processing and QC?

RQ4. How can we minimize uncertainty and improve overall accuracy and results?

2.3.3 DATA AND METHODOLOGY

A systematic literature review, or SLR, is an outstanding tool used often for identifying and assessing relevant studies to a specific problem or a fascinating question. Literature review is about four basic questions, which are highlighted in the section of the research questions (RQ). The SLR was conducted using the Scopus literary database, a valuable source widely used by scholars around the globe, including the international participants of the world’s oil business.

For the analysis of data collected from Scopus collection, VOS viewer software was used which is used for constructing and mapping Network Analysis diagrams based on bibliometric information (Peterson & Cavanaugh, 2020). Again, to execute the systematic literature review, a detailed review process was developed. This approach served as a checklist to ensure the review was done effectively in a manner that is comprehensive and across board without any chance of bias or mistake. Following structural approach was used to carry out the review process as shown in the figure 10:

Figure 10: Flow chart for Systematic Literature Review

2.3.4 SEARCH STRATEGY & METHODOLOGY

Search strategy was one of the first key steps in Systematic Literature review process, where following key parameters were established: -

- Identify the key sources of data (e.g. list of databases) to be searched
- Formulate the strategy of keywords to be used to fetch the most relevant results for the review.

SRL exercise concerned with Geosciences, Petroleum Engineering, and Geology domains, and sought to investigate the application of AI and ML approaches for petrophysical well log construction. Thus, database search was conducted thrice in SCOPUS, Web of Science & Google Scholar covering Industry/Academic Journals, Magazines, Dissertations/Theses, Books and Conference proceedings from all the providers. Consequently, the articles were thoroughly reviewed in order to determine the study's relevance with the laid down objectives. Using keywords from 2003 – 2023 and their permutations and combinations, articles totalling 1177 were obtained and subjected to VOS Viewer analysis. While AI/ML methods began making inroads into the Exploration and Production (E&P) industry around the early 2000s, their widespread acceptance and popularity have surged notably in the last five years. To ensure

a comprehensive literature review, I have selected a review timeframe spanning the past two decades, from 2003 to 2023 as shown in the figure 11.

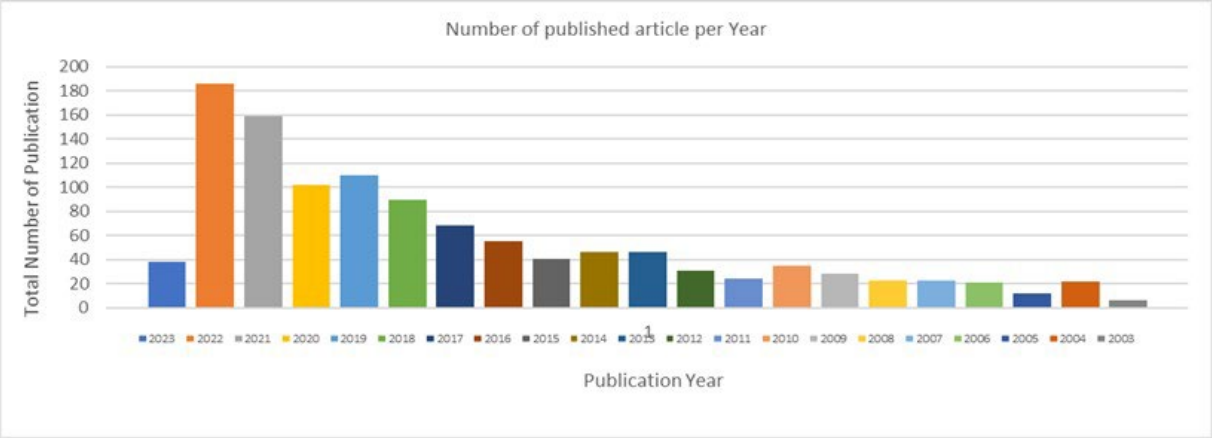


Figure 11: Yearly distribution of published articles

An important tactical manoeuvre used during the literature review was to determine proper words to use when searching through the papers in consideration since the different terms are often used interchangeably or in combination including their acronyms such as AI, ML, Log reconstruction and Log estimation. To overcome this, different forms of both the keywords, as well as their abbreviations, were used for the article searches (Table 6).

Table 6: Various combinations

Key Word Selection and Refinement Criteria	
Criteria – 1	Well log estimation using AI
Criteria – 2	well log estimation using ML
Criteria – 3	well log estimation using machine learning
Criteria – 4	well log estimation using artificial intelligence
Criteria – 5	well log reconstruction using machine learning
Criteria – 6	well log reconstruction using artificial intelligence
Criteria – 7	well log reconstruction using AI
Criteria – 8	well log reconstruction using ML

2.3.5 INCLUSION AND EXCLUSION PROCEDURE FOR ARTICLE SELECTION

Scanning the above computerized internet resources and archives resulted in the appearance of 1177 studies. Specific criteria have been elaborated to include/exclude study papers based on

the PRISMA criterion set up for study reviews to select the most relevant papers and to prevent bias in such papers. The definition of an acceptable factor is:

(1) Articles in English language only.

(2) Articles meeting keyword search criteria defined in table 2 only.

(3) Articles that can respond to the study queries posed in section RQ.

Excluded were documents with duplicates, irrelevant content, abstract-only publications, and studies that could not be compared to study topics. Such papers, parts of the research, were published in scholarly journals, and, papers presented at key international and national conferences. Therefore, based on these inclusion and exclusion criteria, a VOS viewer analysis was done to discover dependencies and relationships. The Co-occurrence analysis of the study was done in VOS viewer by using the results of the literature available and the articles fetched using the above keywords from SCOPUS, Web of Science & Google Scholar database.

2.3.6 RESULTS & DISCUSSION

Co-occurrence Analysis

The co-occurrence maps in the study employed the results of available literature retrieved by entering the identified keywords in the SCOPUS, Web of Science & Google Scholar database to visualize in VOS viewer analysis. Analysis of the kind of literature available through the different fetched research works using the keywords were of interest to VOS viewer. The analysis was designed to use the frequency to examine the extent to which the words were related within the study documents. Given the research aims, the I decided to concentrate on the trends of well logging, well logs, and AI/ML together so as not to deviate from the set objective. There were no exact matches to the term “reconstruction,” but the term “estimation” was suitable as a keyword. From six thousand seven hundred and sixty-seven total keywords, using the cutoff of fifty repetitions for any given term yielded an overall count of fifty-eight total keywords with co-occurrence patterns.

The total weight of the object was therefore selected as an incidence plan in this study, defining the size of the mark and circular boundaries of an entity. The label and circular of a commodity extend in size in inverse ratio to its frequency. The group in which a thing corresponds to has influence on the colour possession of the object. Connections are represented by lines connecting two objects across them. The separation that is put between ‘’two’’ elements in the

network representation, to some extent gives the degree of association they have got in terms of co-occurrence relation (Figure 12-17). Thus, using lines the author represents overlapping and synergy of the dependencies between the things which are the most.

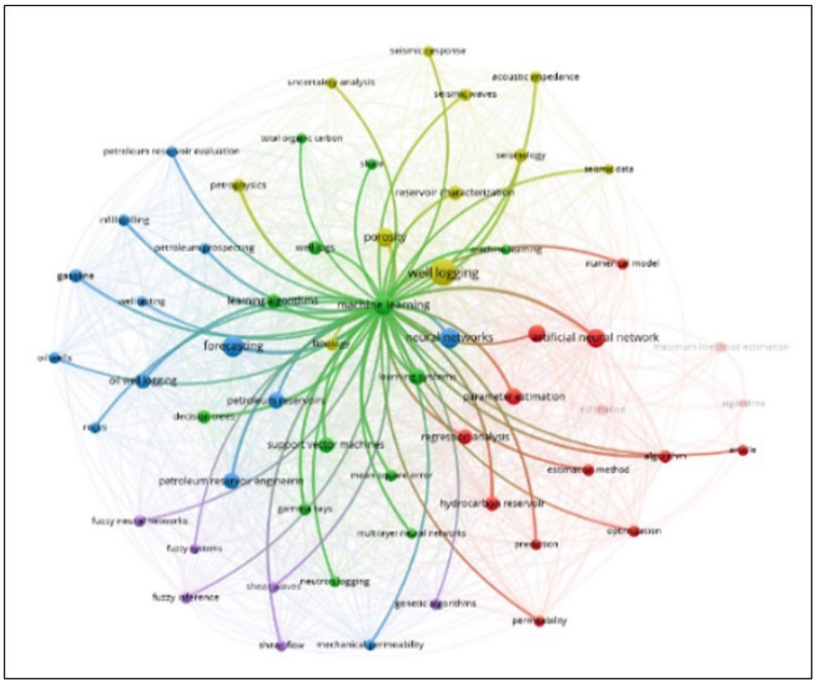


Figure 12: Co-occurrence of various index keywords (Machine Learning)

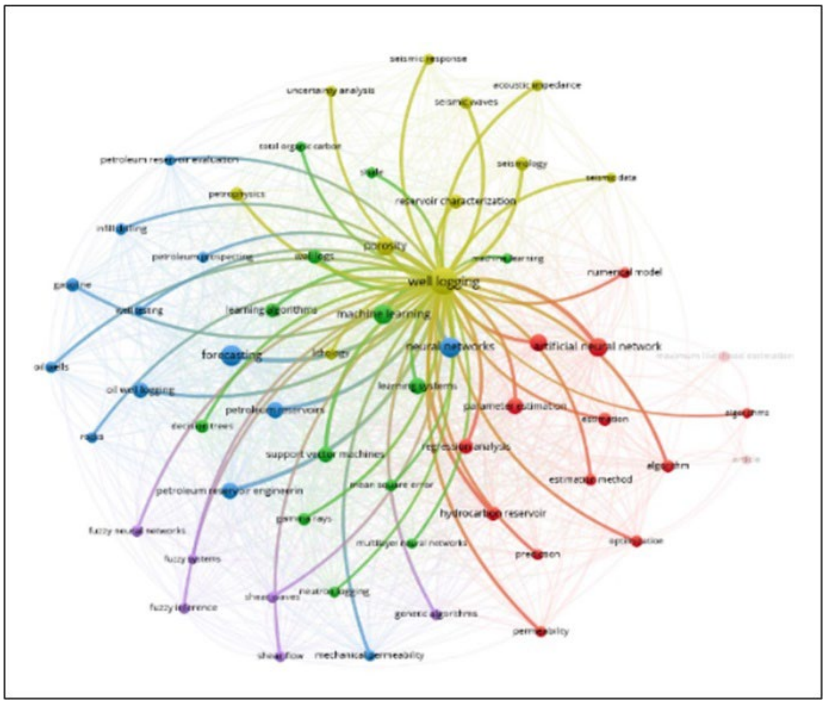


Figure 13: Co-occurrence of various index keywords (Well Logging)

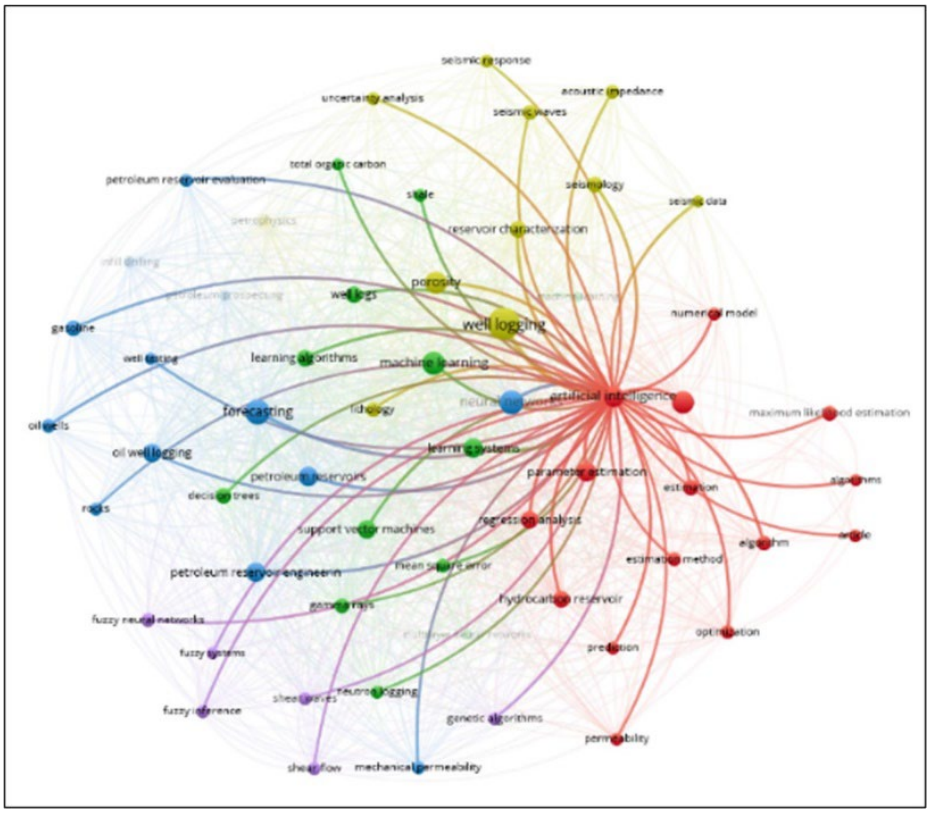


Figure 14: Co-occurrence of various index keywords (Artificial Intelligence)

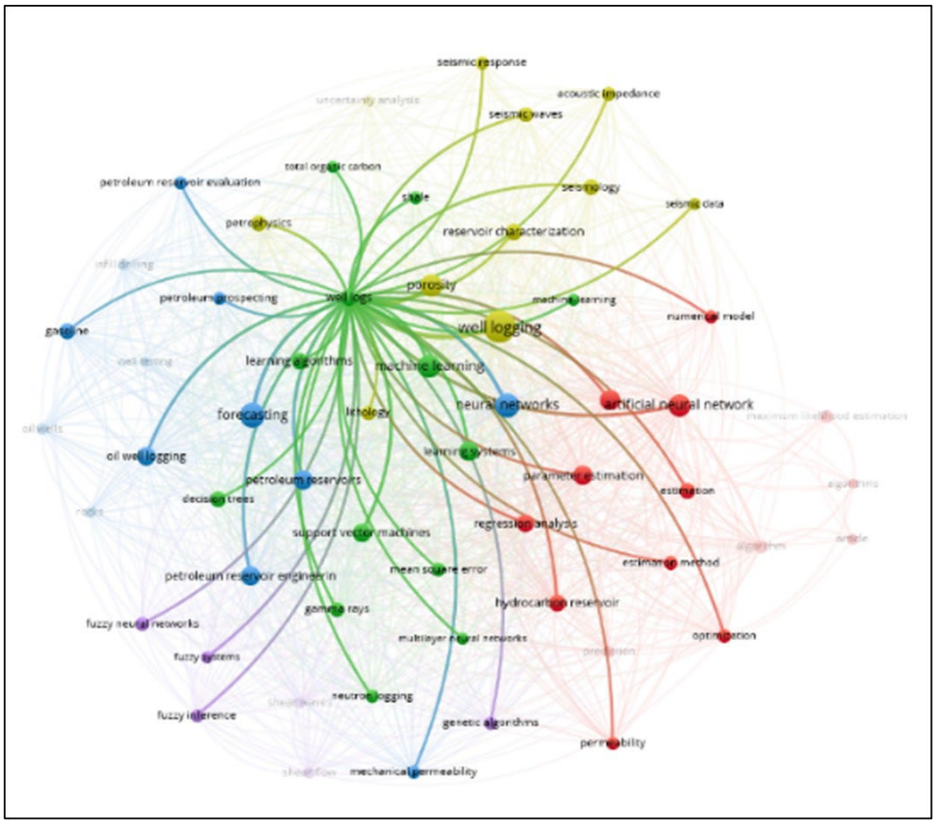


Figure 15: Co-occurrence of various index keywords (Well Logs)

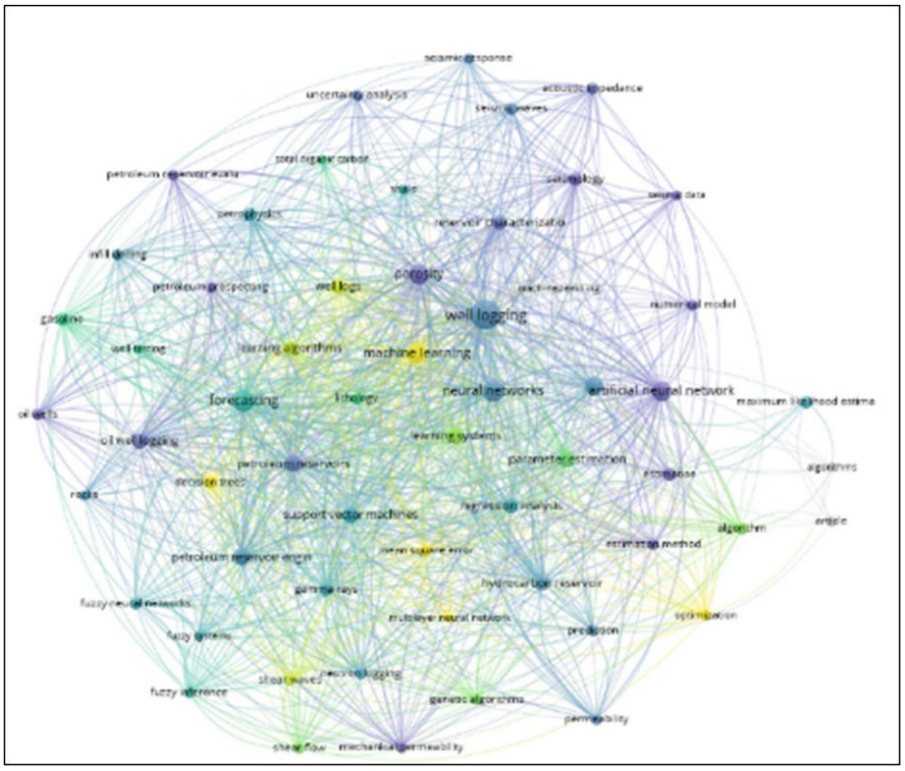


Figure 16: Co-occurrence of various index keywords (Published Year)

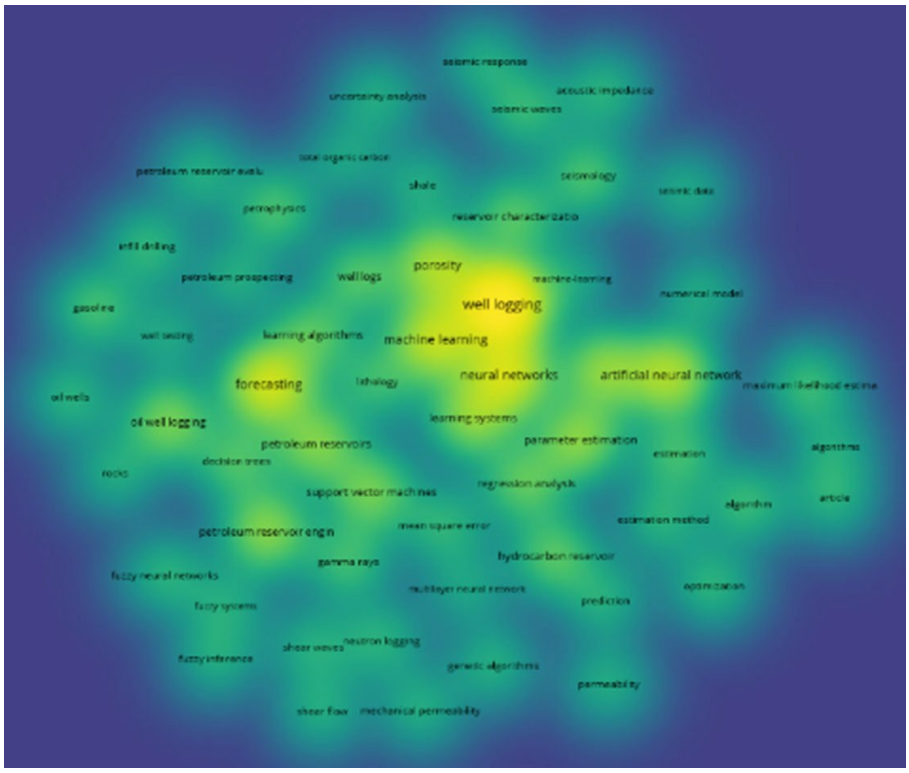


Figure 17: Co-occurrence of various index keywords (total co-occurrence)

Fig 15 – 20 illustrates the network of VOS viewer where circle and a description represent co-occurrence items. As shown in Fig 8, the trends of publications are depicted over the past four years (2018-2021), and Fig 9 represents the occurrence rate of a specific phrase.

RQ – 1 & RQ – 2

From Figure 4, 5 & 7, I can identify the answers to RQ -1 and partially RQ – 2 as explained below. Fig 4, 5 & 7 showing that ML well Logs/well Logging have higher density nearer to each other indicating that they frequently occur together along with Porosity, permeability, estimation AI & NN, regression analysis & Fuzzy NN etc Fig 6 showing AI and well Logs/well Logging have high frequency occurring together along with ML. Based on density and availability of network in Figure 4,5 & 7 it can be observed that Well Logs has been used along with conventional ML methods (e.g. ANN, Regression Analysis, Fuzzy Neural Networks) to estimate the following:

1. Permeability & porosity (Petrophysical property Modelling and Forecasting)
2. Lithology also referred to as Facies or Rock-typing
3. Another value called acoustic impedance or some other property of seismic rocks

The concept summary of the overall cooccurrence analysis done through VOSViewer can be presented in table, which reveals that with the help of using Well Logs, different variety of applications is through ML based methods to estimate porosity, Lithology, TOC. While AI base method has been applied for various types of estimations, predictions & optimizations, the scope is enormous for estimation/prediction of various well logs where the present work is scarce (Figure 8 & 9).

Table 7: Summary of Co-occurrence analysis via VOSViewer

Method	Applications	Method	Applications	Method	Applications
Machine Learning		Well Logging	Porosity	Artificial Intelligence	
			well logs		Numerical Model
	Porosity		reservoir characterization		Maximum Likelihood estimation

	Well Logs		Petrophysics		estimation
	Well Logging		uncertainty analysis		regression analysis
	Neural Networks		seismic response		well logging
	Learning Systems		seismic waves		prediction
	Lithology		acoustic response		optimization
	TOC		Seismology		porosity
			Lithology		permeability
					parameter estimation

Less modern AI applications for log estimation are shown in table 7, as well as the amount of research related to missing well logs (RQ-1). In turn, RQ-2 demonstrates that the greatest amount of research on predicting petrophysical properties with the use of well logs and related methods and their advantages and limitations has been published only within the last five years.

Artificial neural networks (ANN) (Parapuram et al. 2017)

The genetic algorithm (Rahmanifard et al. 2018; Rajabi et al. 2013)

Least squares-support vector regression (LS-SVR) (Venna et al. 2018; Zhao et al. 2015)

Extreme machine learning (ELM) algorithm (Sayyafzadeh et al. 2016; Bishop 1995)

Fuzzy logic (FL) (Karimpouli et al. 2015)

Random Forest (RF)

Bayesian network (BN)

Table 8: Summary of the AI & ML methods via VOSviewer network analysis along with advantages/limitation to be used with well logs

Algorithms	Description	Application	Advantage	Limitations
Artificial neural networks (ANN)	Mathematical design called ANN was developed by biological neural structures, designed to emulate their information processing abilities. These networks learn from input data, identify patterns, and subsequently make predictions based on the trained model (Auda et al. 1999; Ayala et al. 2005)	Regression	Basic algorithms to learn It might be preferable to any other algorithm using the information at hand. Never depends on any function's linearity	They are Blackbox in nature hence hard to understand or interpret Lack the ability of generalization as tend to memories data
The genetic algorithm	Genetic algorithms (GAs) mimic natural evolution, making them suitable for multi-objective optimization tasks and adept at resolving conflicting objectives, providing robust solutions in scenarios with multiple feasible solutions. (Velez-Langs 2005)		Although GA is very effective and simple to use, its operations are uniformly unpredictable, inconsistent, extremely fluctuating, and non-differentiable.	Limiting the quantity of data sets, additional sets of data will produce more precise findings.

Least squares-support vector regression (LS-SVR)	LS-SVR comprises supervised learning methods that analyse data, recognize patterns, and solve linear equations for classification and regression analysis, avoiding convex quadratic programming.	Regression	The LS-SVR provides results with higher accuracy. It handles noise efficiently and over fitting occurrences is less.	Since the LS-SVR system uses the binary categorization method, it must take a long time to cross-validate the variables.
Extreme machine learning (ELM) algorithm	A extended one-layer neural network with feed forward processing that uses a random set of parameters across both input and output stages is the foundation of ELM.		This method may be utilized for issues where finding a solution is difficult or impractical.	Smaller datasets does not provide accurate results
Fuzzy Logic	Instead of the typical true or false method that today's machine uses, fuzzy reasoning is an algorithm-based strategy for computation focused on levels of significance to the real data being input.	Classification, clustering	Quickly, easily, powerfully, and with just small effects on the natural world Provides a representation of networks that combines	If a computational framework already exists, FL uses it only when computer power is limited. Due to the structure's absence of

			numbers and symbols.	arithmetic it is difficult to verify its qualities.
Random Forest (RF)	Random Forest (RF) is a group learning approach that, while being trained for regression, sorting, and other tasks, creates a large number of trees of choices.	classification, regression, clustering	Resilient to data noise May succeed in learning	Poor outcome with training information connected to attributes

RQ – 3 (What are various considerations like pre-processing and QC) and RQ – 4 (How can we minimize uncertainty and improve overall accuracy and results)

Log data, integral to oil companies' on-going well programs, spans several years with variable quality and origins affected by borehole conditions, logging tools, mud type, and processing parameters. Essential petrophysical modelling and interpretation demand rigorous log quality control, with initial processing often handled by acquisition services companies. Work performed by (Gerges et al. 2022; Akkurt et al. 2018; Liang et al. 2019; Singh et al. 2020) highlights the prominence of AI/ML based log QC to be considered in the reconstruction.

Domain Input:

Each ML algorithm using drilling data at a field dimension must take into account topographical and geographical information since it could prove to be just as useful as well statistics.

ML algorithms surpass human data processing in efficiency and accuracy but still rely on expert evaluation for parameter selection and quality control, necessitating user-interaction tools to ensure transparent and efficient workflow management.

Pre-processing Input:

- Data collection and harmonization for mnemonics and Units
- Identification of correct log interval for reconstruction
- Identification of correct set of log responses to be used for Artificial Intelligence and Machine Learning based training for log QC and reconstruction.

Pre-processing of data, conducted prior to AI/ML training, includes handling bad hole intervals, constant values (usually at start and stop depth intervals), and linearizing resistivity logs to ensure compatibility with other log data for tasks such as Log Normalization and clustering.

Outlier Detection:

Anomaly detection, as the process which is aimed at the identification of outliers is one of the critical stages of the development and integration of AI and ML model for log reconstruction. According to Gerges et al. 2022; Singh et al. 2020, Density-Based Local Outlier Factor (LOF) which is an unverified technique for outlier detection (Breunig et al. 2000) performs better than one-class support vector machine (SVM) and isolation forests (IF) outlier detection algorithm. LOF computes a score for anomalies for any given point of data is based on how deviant the density of the given point is to the density of its neighboring points. This property as stated by (Han et al. 2011) to be true provided that the example in question has an anomalous score, above a given threshold or on a different note, its regional density preferably lower than that of near instances. Unlike a primitive way of assigning every observation as an inlier or an outlier, LOF presents a level of oddity and stands in contrast to some other identification techniques of outliers. Such a level of oddity is referred to as the Local Outlier Factor (LOF).

The last intermediate variable in result uncertainty and accuracy means the uncertainty about outcomes of decision making process and the degree of accuracy that the decision makers are able to achieve in decision making.

There is a lot of emphasis needed in this area to get the output of AI & ML based Log reconstruction as low as possible on uncertainty and as high as possible on confidence and accuracy which will be determined based on sigma (σ)factor between reconstructed (predicted) logs adopted from blind wells. Sigma (σ) relates to how variable a specific piece of data is: either the information points touched have high correlation or low density.

2.4 LOG RECONSTRUCTION TECHNIQUES

Log reconstruction methods are critical in the oil and gas industry due to the problems associated with missing or incomplete well log data. The conflict of several data gaps is that there is general methods available which allows the analysis and characterization of the reservoir without larger disruptions. The traditional methods include linear interpolation & extrapolation in which the missing data points are estimated using adjacent log data. These methods appear quite direct although modeling the log measurements are based on the simplistic assumption that the limitation between the values of the logged measurements have a linear constraint which may not be practical when dealing with the intricacies of geological structures. Others include multiple imputation whereby missing values are imputed by developing statistical models that take into account the association between the missing log types, and regression techniques that use the known log types to estimate the missing log type. For example, since the targets in regression models are numerical, it can be built using complete logs' data in order to derive the missing log values by adopting the other available logs. Also, spatial statistical procedures such as kriging offers statistically correlated estimates given the spatial relations in the data and therefore appropriate for location with spatial dependency.

In recent past, the scope of machine learning and artificial intelligence have come as important tools in the reconstruction of logs. ANNs have been applied in the past to model well log data due to their capacity to model non-linear relationships and SVMs and decision trees also show potential for this technique. These methods can thus use the previous data patterns and could bring out a more useful estimates of the missing values. However, these techniques are dependent on the quality and the amount of training data available at the disposal of the modeler. However, most of the existing approaches suffer from few problems, such as dependence on full data for training, noise sensitivity, and inapplicability in different geological settings. Moreover, the applicability of all these techniques under a single umbrella for performing log reconstruction in end-to-end fashion has not been extensively studied. Mitigation of these limitations will be crucial for improvement of current accuracy and reliability of reservoir characterization, especially when it comes to the proposed Integrated Approach on AI/ML Assisted Log Reconstruction Methodology for Estimation of the Missing/Non-Recorded Logs. This new methodology is suggested to develop further on the current long-standing methodologies also taking into consideration its demerits, thereby offering a improved technique for the handling of possible missing log data.

2.5 WELL LOG RECONSTRUCTION USING ML

The prediction of lithofacies (Zhao, 2021) and fluids is fundamental for defining subsurface geological reservoirs and has profound importance in expounding subsurface resources and reservoirs. Normally these predictions involved using well-log data and recently seismic data has been incorporated for better and more precise results. XGBoost has become quite famous as a machine learning approach because it is very useful when handling data with mixed types of data including seismic and well-log data. Compared with other algorithms, XGBoost has shown high prediction accuracy in lithofacies by Shaheen et al. (2021) because of its ability to use weak learners to construct cumulative models especially when applied to high dimension data. Subsequently, feature engineering is useful for enhancing lithofacies predictions significantly. By properly filtering of seismic amplitude, frequency, and well-log attributes like gamma ray, resistivity, and porosity, the models can provide better discrimination of both lithofacies and fluids. For example, when interpreting the two attributes of seismic data and well log data and combining them together, gain is much better sense of location of gas sandstone which is relatively scarce in the data set. However, there are some limitations when it comes to formulating a balanced dataset and, hence integrating geological continuity. Restrictions on spatial data are typically used in these models to guarantee the continuity of predictions, critical in lithofacies with different spatial densities. Aid strategies such as SMOTE (Synthetic Minority Over-sampling Technique) helps in overcoming class imbalance that thrives within the rare types of lithofacies like gas sandstone to improve the machine learning algorithm capability of learning from the classes. However, there are some drawbacks for XGBoost based methods. The primary limitations are their dependence on data quality and the need to make predictions based on a set of balanced data; therefore, making the predictions inaccurate or overfitted in regions with poor or unavailable well data ensure log data affinity which boosts prediction uncertainty in geologically challenging terrains where well data is unreliable.

A recent advancement in the application of ML (Jaya, et al., 2021) is the ability of making seismic predictions for rock properties in areas that lack enough well-log data sets. Classic approaches are mainly based on heuristic dependencies between seismic attributes and rock parameters and, therefore, might be represented by approximate mathematical models only. Other methods include Support Vector Machines (SVM), Random Forest (RF), and Artificial Neural Networks (ANN) that were used to reduce the disparity between little well log data and big seismic data. These work by learning the complex patterns and the interactions from

training sets; they therefore generate predictive models that surpass conventional statistical models. However, this performance is disappointing where well data is lacking or discontinuous and this is a common occurrence in operations in the field.

Advanced Feature Selection along with Feature Augmentation for Improved Log Prediction: Here selection proves crucial for all the well log prediction models in the machine learning system. Filter inputs, such as seismic attributes, can be especially importance for establishing the model's accuracy. Despite these advances, the performance of ML based approaches is normally reduced due to inadequate training data. More severe is a lack of sufficient well-log depth coverage which implies poor generalization and therefore, prediction error. The proposed integrated approach on AI/ML assisted log reconstruction methodology framework avoids this limitation by including automatic well log regularization (ALR), which improves accuracy by 30% with 28% less training data.

The proposed thesis framework leverages a novel feature augmentation strategy integrated into the ALR workflow, allowing the model to exploit intermediately predicted logs as additional inputs, which significantly enhances the prediction of elastic logs. Standard machine learning workflows for well-log prediction involve training models using available logs from a limited number of wells, followed by applying these models to blind wells (wells without complete logs). Several approaches have utilized workflows that incorporate multiple stages of prediction, optimization, and validation to improve accuracy. However, these approaches typically depend on the availability of a dense dataset and are prone to overfitting when the training data is limited. Additionally, they often require extensive manual intervention for hyperparameter tuning and feature engineering, which can be time-consuming. The proposed AI/ML Assisted Log Reconstruction Methodology introduces an automatic and iterative workflow that not only predicts missing logs but also optimizes the prediction process by using augmented features. This eliminates manual intervention and improves overall efficiency, addressing the drawbacks of existing systems.

Well logging incorporating ML (Kim, Choi, Kim, & Byun, 2020) has recently become increasingly popular since it helps to identify and improve the interpretation of subsurface geological formations interlayered within a well. Namely, conventional approaches apply substantial assumptions, empirical correlations, and idealized models, and thus may fail in quantitative determination of the rock parameters. The current research has shown the applicability of ANNs and deep learning modalities for pattern extraction from the log data.

For example, Chen et al. classified the mineral compositions using a multilayer perceptron (MLP) with conventional logging data, which obtain higher accuracy than the regression algorithm. This approach successfully capitalized on nonlinearity found in the geological data; however, they were frequently suffering from overfitting mostly in cases with small or unbalanced training data. Therefore, when using ANNs decision makers will benefit from improved predictive power, however simultaneously they will have to lose some reconstructive aspects of the model, complexity of the analysed problem and insufficient amounts of data can hamper performance of the network.

Specifically, well-log analyses by using machine learning (ML) (Sharma, Burak, & Nygaard, 2023) have received considerable interest in the last few years due to its enhanced predictive tools for subsurface earth properties. Measurements like bulk density, porosity and other aspects of rock properties are made from well logs but sensor offset remains an issue. LWD sensors are typically mounted several meters behind the drill bit and therefore the logging data is delayed and often only partial. The failure of the LWD sensor to provide data in real time for comparison with drilling data is due to this offset. This problem, which is nevertheless rather common in literature, has been addressed the same way by several researchers: trying to overcome it through the use of a number of machine learning algorithms to predict the missing or the delayed data at the bit by depending on petrophysical and drilling parameters. These models make use of data from across several wells, and various categories of machine learning algorithms including (MLR), (KNN), (RFR), (SVM) and (ANN), were tested for their prognosis capabilities. Traditional methods often rely heavily on empirical relationships and simplified models, which can lead to inaccuracies in predicting rock properties. Recent studies have demonstrated the utility of artificial neural networks (ANNs) and deep learning models in extracting complex patterns from log data. For instance, (Cheng, et al., 2019) utilized a multilayer perceptron (MLP) to predict mineral compositions based on conventional logging data, achieving improved accuracy compared to conventional regression models. This approach successfully leveraged the nonlinear relationships inherent in geological data; however, it often struggled with overfitting, particularly when the training datasets were small or imbalanced. Consequently, while ANNs provide enhanced predictive capabilities, they also require careful consideration of model complexity and data sufficiency, which can limit their effectiveness in certain scenarios.

In an attempt to predict petrophysical properties such as porosity and bulk density at the bit, researchers have employed data from both Gamma Ray (GR) logs and Measurement While

Drilling (MWD) parameters. For instance, K-nearest Neighbour (KNN) regression has shown significant promise in addressing the sensor offset challenge. Using data from multiple wells, the KNN model was able to predict porosity and bulk density at the drill bit with high accuracy, achieving a coefficient of determination (R^2) value of 86% for porosity and 74% for bulk density on a validation dataset. Despite its effectiveness, KNN, along with other models such as Random Forest and Artificial Neural Networks, often suffers from sensitivity to hyperparameters and the need for large datasets to maintain high accuracy in predictions. The drawback lies in the computational complexity and the potential for overfitting, especially when dealing with sparse or noisy data.

In contrast, Support Vector Machine (SVM) regression has been found less effective in this domain, primarily due to its poor performance in managing non-linear relationships in complex well-logging datasets. While SVM is highly effective in classification tasks, its application in continuous regression problems like porosity and bulk density prediction leads to higher mean squared error (MSE) values compared to other models. SVM models also require intensive computational resources for training, making them less suitable for real-time drilling operations. The major limitation of SVM is its sensitivity to parameter tuning and difficulty in handling high-dimensional feature spaces effectively. Despite the progress in using individual machine learning models, the predictive accuracy and real-time efficiency remain suboptimal due to model limitations, noisy data, and complex non-linear relationships in the subsurface. The proposed Integrated Approach on AI/ML Assisted Log Reconstruction Methodology for Estimation of Missing/Non-Recorded Logs thesis framework seeks to address these issues by combining the strengths of multiple models. The integration of AI/ML models in a hybrid approach can reduce overfitting, enhance the handling of non-linear data relationships, and provide a more robust solution for real-time drilling analysis. This approach will ensure better accuracy and speed in predicting petrophysical properties, overcoming the limitations of individual models.

For the last several years, various Machine Learning (ML) (Jia, Zong, Qin, & Lan, 2023) approaches for the estimation of direction and magnitude of hydrocarbon source rocks are found to be quite successful by utilizing well log data as well as results of seismic inversion. Source rocks containing hydrocarbons are essential in the production of hydrocarbons, and distribution of source rocks is crucial to the generation of hydrocarbons. A primary means for evaluating these source rocks is the Total Organic Carbon (TOC), which gives the amount of organic matter present in rock and is an essential check for hydrocarbon prospects. By

predicting TOC, the researchers want to predict where those source rocks can be found within the hydrocarbon-bearing basins so increasing the location of well placement. This approach involves partitioning of hydrocarbon source rock quality based on TOC values where the rocks may be described as normal, medium, good, and excellent. In an attempt to improve the reliability of these prognoses, qualitative parameters that controls hydrocarbon generating source rocks are identified utilizing techniques of cross-plot and correlation analysis. Machine learning techniques are then applied to forecast the areal distribution of hydrocarbon source rocks based on an association between these parameters and source rock quality. To assess the ability of the ML models, several models have been tried including Random Forest, GBDT, XGBoost, and HGBDT.

In reference to comparative analysis, HGBDT again performs slightly better in terms of accuracy in spatial distribution of hydrocarbon source rocks. The efficiency of these models is usually measured by confusion matrix and ROC curve; for HGBDT model the results are almost perfect both in comparison with the well data and the depositional properties of the region under investigation. That these machine learning methods can be deemed successful provide evidence on the importance of ML in making the needed subsurface predictions, particularly for hydrocarbon finds. However, the weakness of this approach is in its strong data demands, specifically large quantity and high quality. These can be scarce or noisy, filled with meaningless information or managed with missing data. Furthermore, even specific models which work well are, for instance, HGBDT, could face the challenge of dealing with non-linearity and sparsity of the signals in the well logs and seismic inversion results. To counter these restrictions, the Integrated Approach on AI/ML Assisted Log Reconstruction Methodology for Estimation of Missing/Non-Recorded Logs thesis framework will fuse the models to improve AI and ML based approaches as a solution to noisy and incomplete data. This integrated approach will improve the reliability of the TOC predictions, hence improving the reliability of spatial distribution predictions of the hydrocarbon source rocks.

Hydrocarbon reservoir assessment (Esfandyari, 2022) by petrophysical properties are critical in defining their potential and finding the best ways of production. Specifically, the evaluation of well logs including density, resistivity and porosity well logs is very valuable for gaining information in respect to the geology and the type of fluids within the reservoirs. Previous research has demonstrated that conventional approaches for estimating the petrophysical parameters such as shale volume, effective porosity and water saturation involve simple mathematical relationship and cross-plotting which, in some cases, proposed equation can lead

to approximation and uncertainties because of geomodelling of subsurface formations. For instance when assessing the Mid Cretaceous rocks in the oil field a range of log data was used to establish the rock and fluid volumetric indices, which demonstrated variability between lithologies of the field with regards to parameters such as total porosity and water saturation. However, the mentioned empirical methods can give low accuracy in the case when thin beds or heterogeneous formations evaluation is needed, or data volume is insufficient, as they lack proper consideration of spatial correlations between petrophysical parameters. The proposed method attempts to overcome these drawbacks using optimum segmentations and heuristic AI/ML algorithms to reconstruct missing log data and improve the robustness of the Petro physical analyses.

Advanced Techniques in Reservoir Characterization: The characterization of reservoirs, especially in formations such as the Sarvak, is not only about determining the petrophysical properties of an interest but also about understanding the general geological characteristics of the formation including fractures and lithological changes. In traditional methods, the thickness and other parameters of reservoirs are estimated through measurements from well logs and the interpretations and personal biases and opinions of geoscientists play a crucial role. Recent development has integrated machine learning algorithms into the interpretation process of well log data more objectively and accurately. For, example, the application of ML can be in developing models that use logs to estimate petrophysical parameters and compare them with geological ones for further interpretation of reservoir attributes. However, there are similar challenges in the current state of the art of utilizing different forms of ML in reservoir characterization, such as (a) data quality and (b) training data requirement. This means that when the log data is scarce or partial there are risks of overemphasising certain formations through overfitting than is actually the case or underestimating significant geological characteristics. The proposed thesis framework presents a novel way of reconstructing well log data for the purpose of making it more suitable for the training of the ML models, and thereby more efficient reservoir characterization.

Addressing Variability in Reservoir Properties: Heterogeneities in petrophysical properties in a reservoir can therefore influence its assessment in terms of evaluation as well as in management. In the Sarvak Formation, based on research, it has been realized that component aspects like water saturation (Sw) can greatly fluctuate, and such oscillations depend on aspects like fracturing as well as changes in lithology within the formation. The variations in the permeability and productivity characteristics may not be well captured by using the more

conventional cross-plotting and empirical correlations methodologies hence generalizing the reservoir analysis instead of giving a true picture of the current reservoir. For instance, variation in the volume of shale and the level of porosity has a large impact on the difference between the estimated saturation of hydrocarbons and production rates. Additionally, these conventional methods depend on well log data which contain noise and uncertainty, resulting in lower reliability. It is in light of these challenges that the Integrated Approach On AI/ML Assisted Log Reconstruction Methodology seeks to address through the use of an efficient algorithm which in addition to filling gaps within log data seeks to improve the predictive capabilities of the petrophysical evaluations resulting in better understanding of the variability of reservoir and improvement to the decision-making process in the extraction of hydrocarbons.

Reservoir characterization is important in defining subsurface geology with respect to reservoirs, especially in hydrocarbon provinces such as the Niger Delta. Sophisticated statistical modeling along with geophysics has been adopted comprehensively for delineation of lithofacies and pore fluids. Computerized statistical analysis of cross plots of seismic attributes in bi-variate and tri-variate domains was used in a case study of the “AIB-EX” oil field in the Niger Delta. Through various transformations for example the axis rotation, researchers were able to use such cross plot for discrimination of various lithofacies as well as the pore fluids. The major method of analysis was the reconstruction of geologic lithofacies from the geophysical logs major provided reservoir characterization with the help of rock physics analysis and log inversion. With respect to the positions of reservoir zones, the study encompassed three zones: the first, second and third zones identified as A, B and C respectively.

Several lithofacies; shale, sandy-shale, shaly-sand and sand were recognized within the reservoir through the use of average Gamma Ray (GR) and P-wave velocity or velocity passing (Vp) value intervals. The different lithofacies was assigned into GR and P-wave velocity ranges, which acknowledged the potential of seismic data to characterise large scale sedimentary stratigraphy. Additional analysis continued to tri-variate cross plots of GR, P-wave velocity, and resistivity to better differentiate between fluids, more specifically hydrocarbon and brine. These results suggested that pore fluids could be successfully delineated as being either brine or oil by plotting acoustic impedance against porosity. and elastic impedance versus porosity plots, both of which were effective discriminant tools. The study concluded that crossplotting in various domains allowed for the reliable differentiation of reservoir lithofacies and pore fluids. Furthermore, the findings demonstrated that seismic

inversion could be calibrated using the results to produce a high-resolution seismic lithofacies map, particularly when combined with 3D seismic data.

Despite its success, the limitations of this methodology lie in its reliance on high-resolution seismic data and the computational complexity involved in performing the necessary transformations and inversions. In cases where seismic data quality is poor or incomplete, or where data is noisy, this approach may not be able to produce reliable predictions. The Integrated Approach on AI/ML Assisted Log Reconstruction Methodology for Estimation of Missing/Non-Recorded Logs thesis framework will address these issues by incorporating AI and ML algorithms to handle noisy, incomplete, or lower-resolution datasets. This integrated system would enhance the robustness of lithofacies mapping and fluid discrimination by automatically reconstructing missing logs and improving the accuracy of seismic data interpretations.

Determination of quantitative mineralogy (Predicting mineralogical composition in unconventional formations using machine learning and well logging data | IPTC international petroleum technology conference | OnePetro, 2024) in subsurface formations will always be needful for characterization of the reservoir especially the non-conventional reservoirs. The multi-spectral high-definition spectroscopic logging technics, as those ones, used by Litho Scanner for example, provide an opportunity to obtain detailed mineralogical content. However, these methods are rarely applied in practice due to high cost and mechanism of operations involved. In the more complex or unconventional reservoirs, the problem is exacerbated by the presence of kerogen, as well as the structural heterogeneity and anisotropy of the formations under consideration, which complicate the determination of the mineral composition. To address these issues, application of ML algorithms on well logging and thermal profiling data appears to be an efficient one which is cost effective compared with spectroscopic logging. In this context, the implementation of ML models decreases the dependency on costly tools for reservoir characterization while at the same time amplifying the rate and precision of characterization. This approach has been more appropriate in different reservoirs where conventional techniques impose a great deal of challenge.

But still one of the main drawbacks of this approach is that it rather heavily depends on the quality of data available. Despite its benefits, having few or weak data comes with a significant risk of diminishing returns of machine learning, especially where mineralogy is confined or more than one thermal profile exists. Furthermore, the approach could have challenges in areas

of extreme heterogeneity as this normally presents difficulty to the algorithms since ML generalizes and does not always identify the detailed variability in the reservoir. This work aims to solve these challenges by proposing an all-encompassing AI/ML system for the IA-MLARM MLog Estimation Methodology that can provide solutions to missing/non-recorded logs alongside an ability to autonomously develop machine learning algorithms to learn from complex geological environments. This integrated approach leads to more reliable and accurate prediction and better characterization of the mineralogy of the unconventional reservoirs without having to use highly expensive, high-definition spectroscopic logging tools.

Therefore, cross-sectional (Inoue, 2023) comparisons are often restricted by practical difficulties and differences in time observation methods. In recent years many AI/ML based techniques have come into focus for estimating log for the purpose of well logging when the adequate amount of well log data is not available. A major drawback, however, of the application of any ML technique is in its inability to determine outliers accurately and these are important regions in any reservoir especially aquifers and regions that have been affected with historical water injection. To solve this problem, the proposed approach was the integration of fluid substitution with ML using LightGBM to predict the elastic logs on a grand offshore oil reservoir at Abu Dhabi.

This workflow included two ML steps: first, the evaluation of the original elastic logs of the porous medium in situ, and secondly, the elastic properties under conditions free from fluid, based on saturation log and history matching the reservoir simulation model. This study also showed that by implementing fluid substitution and machine learning, the elastic properties of the brine and oil-saturated areas were accurately modeled, according to blind well tests. When plotting predicted P-impedance against Vp/Vs, it was ascertained that the estimated elastic properties were suitable for other carbonate reservoirs. In this approach, I was able to successfully predict the elastic logs for 65 wells under dry and reservoir settings having good correlation to rock physics analysis and giving better estimation than empirical equations/MIME Modelling of the Xu-Payne type. This approach has the following limitations: Since it involves fluid substitution models, the accuracy of the model plays an important role, and the model at relatively low frequencies is used, if it is available. Furthermore, there is a possibility that the application of ML techniques will be problematic in incorporating complicated features of geological field or inadequate data. Due to the the challenges of missing/non-recorded logs, discordant data reconstruction capability using the integrated approach on AI/ML assisted log reconstruction methodology for estimation of missing/non-

recorded logs framework improves on the outliers and complex formations' prediction where incomplete or noisy datasets are encountered.

In recent years, borehole image logs (Howard, Tracey, Shen, & Zhang, 2021) have become invaluable to determine the attitudes and extent of fracture zones in reservoir sections. These logs are able to detect fractures because the rock matrix has a higher electrical or acoustic response than the fluids in the fractures. Nonetheless, quantification of the width of sub-resolution fractures remains problematic because these tools are typically available with millimetre-per-pixel resolution. To address this, a novel AI workflow has been proposed as using ML segmentation in combination with a deep learning neural network to support the visualization of fractures in the borehole image. This workflow has proven to have the ability to greatly enhance the analysis of fracture widths and most importantly, in situations whereby the features are below the ability of the logging tools to capture.

The procedure starts with the separation of borehole televiewer (BHTV) log images into rock and fluid-filled fractures with the help of a machine-learning segmentation algorithm. This tool employs several image processing filters to analyse patterns in the distribution of fractures and rocks in terms of nearest neighbour behaviour. A regression model is built based on a smaller part of the well and the resulted model is used to predict the whole log generating probability for each pixel of being a fracture or rock. Pixels near the rock-fracture interface or pixels belong to sub-resolution fracture are given intermediate probability values, creating a map of probable fracture distributions. For the enhancement of the borehole images the deep learning aspect is implemented by a Conditional Generative Adversarial Network (cGAN). In this network, the created images are improved through continuous comparison with actual BHTV log images leading to better fracture continuity and resolution. However, this AI based workflow has its own limitations towards the further enhancement of fracture image resolutions. The process is very sensitive to the quality of training data.

Advanced computational methods (Singh, Seol, & Myshakin, 2020), and the rich history of machine learning (ML) application in petroleum engineering and geoscience boast of having modern analytical and interpretative tools such as well logs data. Disposable well logs that contain critical subsurface data have, for long been analysed using physics-based methodology augmented by the human interpretation. Unfortunately, due to the increasing application of ML, most of these processes have now been fully automated to work better and with higher degrees of accuracy. Literature review reveals that supervised learning (SL) based models

including artificial neural networks (ANN), decision trees (DT), and support vector machine (SVM) were reported to give very high accuracy for the estimation of reservoir properties such as porosity and permeability from the well log data. In contrast, the UL techniques including k-means clustering and principal component analysis (PCA) have been used to perform lithology classification and anomaly identification on the well logs. Despite their growing adoption, existing ML-based well log analyses face two significant limitations: first, many studies utilize a finite list of the well logs utilizing no approach for discovering which logs compose the ideal set for prediction; second, analysis within the framework of UL classification yields inferior accuracy in comparison with SL. The challenges associated with well-log interpretation have discouraged SLB's petrophysicist from reconstructing the missing or non-recorded logs by directly applying mathematical and statistical methods on the existing database this way, the well-log construction with accuracy and reliability of the missing or non-recorded logs become possible by optimizing the well-log selection and the ML algorithm used in this proposed Integrated AI/ML-Assisted Log Reconstruction Methodology for Estimation of Missing/Non-Recorded Logs. The findings provide evidence that the given automated workflow can identify the ground truth for the lithology with an 80% highest probability of being right using UL and 90% using SL based on six conventional well logs: vp, vs, ρb, ϕ, Rt, gamma ray (GR).

2.7 WELL LOG RECONSTRUCTION USING DEEP NEURAL NETWORKS

CNN (Das, Pollack, Wollner, & Mukerji, 2019) has demonstrated a great potential in seismic processing and well logging problems, where deep neural networks (DNNs) recorded numerous successes. CNNs have been adopted to convert normal-incidence seismic data into elastic models of the subsurface reservoirs, which effectively offers a new method of well-log analysis. Thus, with training of CNNs using such synthetics full-waveform seismograms as generated by Kennett's reflectivity, researchers are able to predict subsurface properties such as P-wave impedance (***IP***). These seismic models, being restricted by rock-physics principles, help extract essential geophysical information. Several recent studies have shown that CNN-based inversion approach is highly accurate, for instance, inversion of Volve field dataset as depicted in the above figure gives rise to a correlation of 82% of predicted ***IP*** with well log data. Moreover, approximate Bayesian computation has been used to compute the posterior distribution of CNN predictions for improving modelling uncertainty of subsurface systems available. Nonetheless, several issues lies with CNN-based seismic inversion. It is clear that

variations in facies proportions, in rock-physics models, and in seismic source such as phase and frequency of the wavelet may well alter the prediction accuracy. This sensitivity shows that the model is not well generalized to different geological conditions. The five-step AI/ML log reconstruction approach to estimate non-missing logs minimizes the above drawbacks by using a wider range of potential well-log pairs and refining the training algorithms to further fine-tune the predictions under variable geologies, thereby optimizing the ML performances generally.

Using DNNs (Du, Li, Zheng, & Srikumar, 2017) in the context of the same application as in the work of logging anomaly detection based on long short term memory (LSTM), Abroshan involves the application of Long Short-Term Memory (LSTM). An example is DeepLog, which uses LSTM networks for system log data, In that case logs are comparable to sequential data type of natural language. Therefore, DeepLog can separate normal logs by itself and then recognize abnormal logs, which is why individuals can observe that DeepLog has a very high accuracy of identifying anomalous behaviour. DeepLog also adopts the element of time ordering inherent in log data since logs essentially record the chronological state transitions in a system along with corresponding important events. Furthermore, DeepLog can incrementally update its model online, can address newly identified log patterns, and new system behavior. Most importantly it is flexible to newer logs as seen in environments that are constantly changing due to software upgrades or structure alternation. Furthermore, DeepLog builds the entire process flow out of log data and they are able to investigate the root cause of the detected anomaly, which is fundamental in solving system and software problems. However, there are some limitations in DeepLog, for instance, DeepLog is sensitive to the setting of feature fields. One disadvantage arises from its fixed sequence of operations based on which it may not be able to detect more general or novel forms of anomalies in well-log data. Additionally, the degree of labeling of datasets which are also large may hinder further applicability of the model especially in conditions of limited resources. To overcome these challenges, the proposed Integrated AI/ML-Assisted Log Reconstruction Methodology for Estimation of Missing/Non-Recorded Logs tries to utilize a wide range of well log combinations for constructions; it also moderates the learning models to set much less dependency on pre-specified well log sequences which can apparently enhance the accuracy of novelty detection as well as log reconciliation across various operational states.

Shear wave velocity (Vs) (Wang, Cao, & Yuan, 2020) plays a central role in reservoir description and fluid detection to support geomechanics and rock physics practices.

Nonetheless, direct Vs measurements from well logs are mostly expensive, thus leaving incomplete Vs data in traditional well logs. Consequently, in the recent past, advanced techniques, including deep neural network (DNN), specifically recurrent neural networks (RNNs), Such innovative models as Long Short-Term Memory (LSTM) have developed into effective means for predicting Vs based on conventional well logs. One of them is applying LSTM with the help of the PSO approach for adapting the hyperparameters of the model applying adaptive learning strategy. This hybrid model makes it possible to estimate Vs by incorporating other well log data that are more sensitive but have high dimensions reduced using techniques such as grey relational analysis. As for the dependencies between well logs and Vs, LSTM successfully learns nonlinear mappings between them. The adaptive PSO optimization still improves the model's performance by competitively matching log data features to the network structure besides yielding higher predictive accuracy and robustness. However, the limitation of LSTM-PSO hybrid approach is still exist in two aspects, such as selection of sensitive logs and the difficulty of tuning the parameters which may affect the scalability of the hybrid model in large scale scenario. These limitations are overcome in the proposed Integrated AI/ML-Assisted Log Reconstruction Methodology for Estimation of Missing/Non-Recorded Logs through the automation in choosing the optimal well-log combinations and improving the flexibility of deep learning models in order to result in accurate Vs prediction in different geology.

The incorporation of (DNNs) (Tripoppoom, 2019) in reservoir characterization especially with regard to fractures in low permeability shale formations received special consideration because of these methodologies' capabilities on not only increasing the knowledge of the reservoir aspects but also the properties of the formation itself. Such an approach can be the Assisted History Matching (AHM) workflow based on the Neural Network-Markov Chain Monte Carlo (NN-MCMC) algorithm. The general approach extends the capabilities of standard MCMC for quantifying uncertainty and distinguishing between local optima and global optima by using a proxy model, which is the neural network in our case, thus reducing the computational load often encountered in MCMC approaches. This combination enables the computation of several realizations of subsurface properties motivated to address the inherent uncertainty of hydraulic fracture geometries and reservoir properties. However, there is still a problem of sensitivity to the initial parameterization and assumptions pertaining to the fracture behavior that affects the quality of the history matching. The following Integrated AI/ML-Assisted Log Reconstruction Methodology for Estimation of Missing/Non-Recorded Logs mitigates these challenges by

offering a more flexible and generative approach for log reconstruction so that they allow the additional information sources inclusion and enhance the predictive performance under a variety of geological conditions improving reservoir characterization and production prediction.

The use of deep neural networks especially the convolution neural network (CNNs) (Lima, Suriamin, Marfurt, & Pranter, 2019)has presented promising results in diverse areas including the geoscience discipline in which conventional techniques prove slow and imprecise. Comparatively, CNN's efficiency in recognizing images at a par with human-like performance through fresh innovations has been relatively restrained in geoscientific fronts including lithofacies identification in core samples. Machine learning that identifies CNNs for identifying images from cored carbonate rocks has been established for the first time. This is done through transfer learning and training models on millions of images making it predict gender with high accuracy of about 90%. This high level of precision not only improves the lithologic description but brings the standardization to that description process, which may result in considerable time reduction and increased uniformity in lithologic interpretations (Quintero-Bermudez, 2018) (Zhu, et al., 2018). In addition, a new labeled dataset helps expand existing source material for further research in the fields of geology and data science. Although these prospects have been made, there are limitations in the applicability of CNNs to various geological environments, on the requirement of high quality and diverse training data. The proposed Integrated AI/ML-Assisted Log Reconstruction Methodology for Estimation of Missing/Non-Recorded Logs deals with these drawbacks by including a wider range of log data and variable modeling techniques which lead to improved prediction and more holistic evaluation of the reservoir subsurface environment in diverse types of reservoirs.

Application of artificial intelligence including artificial neural network (ANN) (Alizadeh, Maroufi, & Heidarifard, 2018) in marker-controlled evaluation of source rock potential has proven meaningful in petroleum geology. Being available and positively associated with the geochemical properties, wireline data is a useful tool for identification of the source rocks and prediction of their richness. Two latest studies have demonstrated how the ANNs and the ΔLogR method can be applied to estimate the missing logs and TOC content while illustrating the ability of intelligent tendencies to estimate other forms of organic matter. The described methodology is also applicable to modelling of the S2 parameter, which represents the present hydrocarbon potential through wireline data and TOC values. The back-propagation method used to train the network on the 70-points dataset fulfilling the nonlinear associated with such

geochemical assessment. Comparing with results of Rock-Eval pyrolysis, Log versus organic richness and lithology both presented a high dependence and further shown that in each simulation, the ANN outputs resulted in better precision than the ΔLogR values. However, the analysis is limited to few datasets and Lithological contexts which might limit the generalization of the ANN models. These concerns are relayed to the proposed Integrated AI/ML-Assisted Log Reconstruction Methodology for Estimation of Missing/Non-Recorded Logs and this work because it introduces more forms of well-log data and more flexible models into the predictions to improve the generality of the predictions in various geologies, making the estimation of geochemical and depositional attributes of complex reservoirs more robust.

Therefore, well log analysis (Das & Chatterjee, 2018) is central in evaluation of the petrophysical properties of reservoir rocks, and further the fluids they may contain beneficial in hydrocarbon search. To explain well log responses like gamma ray, resistivity, density and neutron logs the recent investigations have shown how it can be interpreted across primarily wells in Krishna-Godavari (K-G) basin. When using traditional log analysis, the scientists detected 14 HC saturated zones as well as the zones with gas saturation detected by very low gamma ray values, high deep resistivity and certain crossover trends of neutron and density logs. Through using P-impedance and S-impedance and LMR cross plots have been of great help in outlining the lithology and fluid type facies and improving the resolutions of the various geologic features. For example, the LMR technique enhances the differentiation of the fluid types and carbonate facies relative to previous techniques; the petrophysical parameters such as shale volume and effective porosity were also compared to the actual core and cutting samples. However, the use of conventional methods may restrict the resolution and accuracy of interpretations especially when data variation is high in such geological terrains. The essence of the proposed Integrated AI/ML-Assisted Log Reconstruction Methodology for Estimation of Missing/Non-Recorded Logs is to build on those and identify the missing logs with machine learning methods to perform improved data fusion and the usage of lessons learned from previous log datasets. The rationale for this approach is to improve the is density of Petrographical assessments and other lithological determinations that are central to decision making in more hydrocarbon exploration and exploitation.

2.8 WELL LOG RECONSTRUCTION USING AI TECHNIQUES

This study aims to analyse the use of artificial intelligence (AI) (Koroteev & Tekic, 2021) in the oil and gas industry specifically in upstream sector as the way resources are discovered and produced is changing at a fast pace. These techniques are particularly effective when it comes

to the well log analysis, to give a thorough interpretation of DNNs and other related geological data. I have shown that utilization of DNNs may enhance the classification accuracy of lithology and identification of fluids up to multiple well log types inputs, which can provide an efficient picture of subsurface conditions. Indeed, the automation of such formal processes through applying DNNs with the help of extensive datasets can reduce the time on decision-making and prevent the high risks of operations, which are inevitably connected with hand interpretation of logs.

Nonetheless, issues of data quality and availability still persist, and the complexity of these models would require experts' implementation. An important requirement for the training of DNNs is high-quality, labeled data; however, data integration and management issues are challenging for many organizations, which often leads to incomplete or inconsistent Log records. Also, the technological know-how to train and sustain these AI systems is still relatively scarce within the trade. To overcome these challenges, there is a proposed methodology that serves as an Integrated AI/ML-Assisted Log Reconstruction Methodology for Estimation of Missing/Non-Recorded Logs. The primary objective of this methodology is to enhance the predictive and operational capability over the subsurface properties for arriving at correct and enhanced decision-making tools for the oil and gas exploration and production processes.

The application of artificial intelligence (AI) (Rajaee, Ebrahimi, & Nourani, 2019) techniques in the modeling of GWL has attracted much attention in the recent past as a way of capturing new ways of simulating and predicting the state of groundwater across different aquifers. To identify the most common AI techniques that are used in the context of GWL forecasting, major findings are presented as the results of the systematic review of 67 journal papers from 2001 to 2018, articles that contribute ANNs, SVM, and fuzzy logic systems. This review is cantered on the type of input data that one uses and the modeling techniques that need to be adopted to improve on the prediction models. Although the work shows that AI methods can be effectively applied to model GWL time series, it also identifies some drawbacks of these approaches, Trail-and-error procedures to diagnosis and develop models, as well as requirement of an expert to fine-tune parameters. In addition, there are many AI models that are not standardized creating a problem because different aquifers have varying performance levels. That is why, we assume that there is a need for the further integration of GWL modelling. To overcome these limitations, the present research introduces an Integrated AI/ML-Assisted Log Reconstruction Methodology for Estimation of the Missing/Non-Recorded Logs that embeds

superior techniques of artificial intelligence, improves the quality of the data, and enables the reconstruction of the incomplete log data. This methodology will enhance the predictability of GWL and address concerns that relate to groundwater resources for sustainable development.

In the petroleum industry (Otchere D. A., Ganat, Gholami, & Ridha, 2021), AI has been integrated into many areas of its operation ranging from exploration to reservoir management operations. Of all the used techniques, Supervised Machine Learning Especially Artificial Neural Networks (ANN) has received significant attention due to its performance in modelling non-linear patterned relationships for high dimensional data sets. The results achieved applying ANNs are considerably better than those of the traditional regression approach in aspects such as predictive accuracy, computation speed, and memory requirement. However, it is observed that ANNs have some disadvantages as well, such as a large amount of data are required to achieve the best result in most of the cases, where data is sparse.

Moreover, novel competitive algorithms such as SVM and RVM have also been developed to be efficient replacements for ANNs, and in many cases, have been demonstrated to provide best results even when the database contains limited data. This trend explains the de-emerging of ANN models and the ever-rising demand for SVM and RVM since data scarcity would have posed a big challenge to ANN models. Moreover, indeed, the combination of the various algorithmic methods can also be more effective in characterizing reservoirs, thus truly looking at the methodology of collecting algorithms and evaluating the performance based on each's strengths can indicate better results in reservoir characterization. However, there is a drawback in using single algorithms which limits the performance of the system. Addressing these challenges is the aim of the Integrated AI/ML-Assisted Log Reconstruction Methodology for Estimation of the Missing/Non-Recorded Logs presented in this thesis, which optimizes the use of different AI approaches to enhance the precise well log predictions even when the data is minimized or not available at all.

Study and description of fully or partially naturally fractured vuggy dolomites (Ye, Jiang, Liu, Wang, & Gu, 2024) are challenging because of their heterogeneity and yet they bear large potential of Hydrocarbon resources. New developments have quasi-integrated AI approaches such as unsupervised clustering techniques to improve the identification of reservoir facies, which successfully classifies the reservoir facies by log response characteristics. Principal Component Analysis (PCA) for dimensionality reduction and noise reduction in the input data increase the fidelity of the subsurface conditions. This is because this has a dual approach,

which identifies reservoir levels with an accuracy of more than 90 percent; more than the basic models such as K-means clustering.

However, as different from humans, these intelligent algorithms provide more applicable and accurate result but they might have problem with applying variability in data and differentiate it especially if the area is less described. The major drawbacks of current AI/ML approaches to log analysis have been discussed above; the proposed Integrated AI/ML-Assisted Log Reconstruction Methodology for Estimation of the Missing/Non-Recorded Logs is to address these shortcomings with a wider framework that incorporates a number of AI strategies. This combined work flow also enhances the certainty of reservoir facies design and coping with missing or non-available log data thus contributing to enhance the tendency for reservoir exploration and exploitation.

Plans to execute more wells along with recent development in artificial intelligence techniques (Urang, Ebong, Akpan, & Akaerue, 2020) have improved the quality and speed of interpretations of reservoir property from limited well log data. Conducted in a Bayesian framework and through the use of multi-layer perceptron neural networks, a great potential can be noted within this field. Two distinct neural networks were developed: one that would calculate porosity using the density log data and another which used both RHOB and water saturation to compile for permeability. These networks were able to produce low mean square errors (MSEs) in predictions, and high R-values, pointing to a high correlation between predicted and target values in as many wells as possible. The final R- values of the porosity network are between 0.83749 to three nine eight seven and for permeability, it is zero point eight to five one two to zero point nine seven six four show this model has an accuracy as well. Although impressive, challenges still exist to truly generalize such models with different types of reservoir conditions and potentially even more insidious is the potential over fitting to very specific datasets. The proposed Integrated Log Reconstruction AI/ML-Assisted Methodology for Estimating Missing or Non-Recorded Logs overcomes these limitations by taking a more ecosystemic approach to incorporate missing data and improve model robustness. Not only it enhances the predictive quality of reservoir properties, but also it helps to make resource exploration and management decisions more conclusively dependable by incorporating uncertainties and log data heterogeneities.

AI techniques (Solanki, et al., 2022) have enormous potential to improve reservoir management and optimization from the Oil & Gas sector. Artificial Intelligence applications are becoming

vital to improve the efficiency of already established infrastructure and at scale thereby leading to better outcomes across several operational areas. In this study, a review of AI applications in the oil and gas sector was carried out across different stages: (1) reservoir rock properties estimation; (2) drilling optimization; and (3) production facilities design. In hindsight and looking at how AI models use wire-line log data and seismic attributes, the accuracy of the prediction is necessary for a robust reservoir development plan. Permeability is commonly required for exploitation of the reservoir flow behaviour with in single-phase fluid modeling and indeed some papers have been given to investigate measurement from chosen well-log dataset, displaying the possible difficulties. Although AI approaches represent an important opportunity in drilling fluid design and drill bit selection, the inconsistency and unavailability of data are challenges that hamper their application such as the danger of over-fitting within specific datasets. This will limit the performance of the model, which is a disadvantage as well since our proposed Integrated AI/ML-Assisted Log Reconstruction Methodology for Estimation of the Missing/Non-Recorded Logs systematically treats all these disadvantages to achieve better results. This method may further increase the predictive capability for reservoir properties, and hence support a more effective hydrocarbon recovery process by correctly associating missing or unrecorded log data.

Permeability is one important parameter (Moussa, Elkatatny, Mahmoud, & Abdulraheem, 2018) which has to be known exactly to understand how hydrocarbon reservoirs behave and control several petroleum engineering problems. Core analysis and well testing, which are the traditional methods to establish permeability, can be expensive and time-consuming. In response to these challenges, scientists have started leveraging artificial neural networks (ANNs) in order to study the mapping from core permeability to well log data. Recent developments have resulted in new empirical correlations that employ deep resistivity (RT), bulk density (RHOB), micro spherical focused resistivity (RSFL), neutron porosity (NPHI) and gamma ray (GR) well log parameters for permeability prediction purposes. The newly engineered implementation of SaDE-ANN and EASR have demonstrated a good result by obtaining accurate levels close to 0.0638 for the mean square error (MSE) and 0.98 for the correlation coefficient (CC). Considering the accuracy of the results, I observe that our SaDE-ANN approach essentially improves upon EASR(noticeably). While these advances have been made, issues still arise in terms of generalizing the models across reservoirs and geological conditions since they are based on a small number of limited datasets with specific core measurements. The proposed Integrated AI/ML-Assisted Log Reconstruction Methodology for

Estimation of the Missing/Non-Recorded Logs attempts to ease these limitations by offering a complete protocol with all missing log data, aiming toward improved reliability and applicability concerning permeability predictions in multiple reservoir situations.

2.9 CRITICAL OBSERVATION

- **Diversity of Approaches**: There are a wide variety of methodologies used in log data analysis published, the literature uses traditional statistical techniques to cutting edge AI models. Despite the transformative progress that has been made with respect to machine learning methods, such as artificial neural networks (ANNs) and support vector machines (SVMs), there is not always consensus on the best method for particular applications. The wide range of outcomes as well as the effect that technique has on outcome might argue in favour of some standardized protocols or guidelines dictating what sorts of methods are most likely to work in different situations.
- **Data Quality and Availability**: One of the recurring themes in literature related to missing or noisy log data. Despite the suggested methods for log reconstruction in most current studies, these methods are not enough and have been developed without a detailed consideration of the challenges related to data collection and quality control. This dependence on quality datasets to train machine learning algorithms raises the question of how practical are these models likely to work in actual scenarios where such data can be scarce or biased.
- **Integration of Techniques**: Studies point out to the gap in integrating different techniques for a general log reconstruction. Existing methods often work in isolation and concentrate on a single technique, without the exploration of possible synergy by combining different techniques. Such fragmentation may constrain the more widespread utility of log reconstruction activities, and reinforces the value of an integrative framework that exploits the complementarity among methods.
- **Site-specificity and generalisability across different geological environments**: The proposed technique shall be generalizable across a variety of geologies. A large majority of the applied research studies are site-specific; any inferences made there cannot be generalized. AI/ML models learned on one particular geography might not generalize equally well to other locales, thus demanding more experiments to validate these methods in other domains.
- **Future Research Directions:** Based on the literature review, there are immense gaps in areas related to new methodologies and potential hybrid models of traditional and

modern techniques. The changing landscape of AI technologies therefore calls for continued research and development in the domain to be in line with the changes. This opens an area for future studies to concentrate on the more robust, flexible, and adaptable log reconstruction methodologies in coping with challenges against incomplete or noisy data.

In this regard, while the immense literature review offers very rich insights into the history of developments in well logging and AI application in the oil and gas industry, it also reflects a lot more integrated and standardized approach in addition to adaptability. Therefore, it becomes an integrated approach on the AI/ML-assisted log reconstruction methodology for the estimation of missing and non-recorded logs by trying to deal with these important issues proposing a framework that enhances the accuracy of log data reconstruction as well as fostering collaboration between different methodologies and practices in the industry.

2.10 GAP OF THE RESEARCH

Despite advancements in wireline logging and application of the techniques of artificial intelligence and machine learning in oil and gas industry, important gaps still exist and would be efficient to fill up the gaps associated with challenges facing the challenges of petrophysical log data. An important gap that is drastically missing from the methodologies to estimate missing or non-recorded logs, which can severely affect the reliability characterizing reservoirs. The approaches currently used are very much based on rather conventional statistical methods or simplifications which do not necessarily take into account the often complicated, non-linear relationships in the geological data. Furthermore, although many AI and ML techniques have been promising in reservoir property models and predictions, few comprehensive studies do exist on an integrated use of such techniques with specific applications towards log reconstruction.

These are mainly fragmented pieces of literature discussing individual topics about reservoir assessment, such as estimation of porosity or permeability or the general need to acquire accurate log data over several parameters at one time. Additionally, relatively less is known about the various types of noise and inconsistencies in the log data that affect the performance of AI/ML models in practice. This means further exploration that is not only on new algorithms but also assessing their robustness in the presence of real logging challenges. The integrated approach on AI/ML assisted log reconstruction methodology for estimation of the missing/non-

recorded logs fills these gaps by presenting a systematic and comprehensive framework to enhance the accuracy and reliability of petrophysical log interpretations.

2.11 INDUSTRIAL TOOLS

Currently, in commercial space, there are multiple tools available some of the most common ones are - Techlog from SLB, Interactive Petrophysics from Geoactive and Geolog from Aspen. In petrophysical interpretation, there are difficulties associated with the complexity of subsurface formations, variability in well log data, and the limitations of measurement techniques. Some of the key challenges include:

- Limited data quality
- Unavailability of complete sets of logs
- Complex interpretations
- Lack of experience and expertise

These solution addresses the common challenge where interpreted results are available but lack traceability to the underlying model and the workflow used for their creation, often necessitating time-consuming manual efforts to reconstruct these models.

Input

- The LAS files can directly be pushed to the AI platfoprm project from petrophysical application.
- Minimal customizations (log names, mnemonics, and variables in the Python script) are required for this workflow.
- All the wells in training must follow the same mnemonics for logs and should have the same units.
- All the wells should have only one dataset in the project.
- Wells should be grouped under Training wells and Testing wells.

Model summary

The available solution from various vendors includes two distinct models. The first model employs the Random Forests algorithm and includes the following features:

- Automated hyperparameter tuning

- Uncertainty computation

The second model introduces a more recent approach, utilizing a neural network architecture known as Long Short-Term Memory (LSTM). This architecture considers the sequential information present in the input well logs, thereby enhancing the ML model's ability to predict output properties at different depths

Output

The workflow yields individual CSV files for each well included in the prediction. These CSV files encompass the entirety of the initial logs, along with the corresponding predicted logs and can be seamlessly visualized within the Techlog wellbore software or any other visualization tool.

Value

Efficiency

Leverages ML to learn the underlying model consistently and objectively, eliminating the need for manual reconstruction.

Automation

With a single click, automates the tuning of model hyperparameters for well log interpretation, saving time and reducing human intervention.

Improved Decision-Making

Enhances the quantification and tracking of uncertainty, enabling better decision-making backed by reliable data

2.12 SWOT ANALYSIS

SWOT analysis is being presented for some of the common petrophysical tools available in the commercial market today. The tools which has been compared are – Techlog, Interactive Petrophysics, Geolog and Petroworks

Strengths: Most of these tools offer varying degree of user interface (menu driven and automatic) for interpretation. The tools offer both conventional and advanced multi-mineral petrophysical analysis. Since these tools are backed by large organizations, they have a solid support and training infrastructure.

Weakness: All these tools are relatively expensive and not possible for small organizations or individuals to afford them. The price per user license may vary from US$ 25,000 to upto few US$ 250,000+. Since the tools are modular, some of these are quite complex to understand. None of them have extensive AI integration.

Opportunities: Since tools are not extensively integrated with AI and new generation methodologies, current study and work opens opportunities for AI & ML based petrophysical interpretation in these tools, which will also help to expand their existing market.

Threats: There are many simple open platform solutions being developed and available which poses threat to the existence of these tools. Fluctuations and up/down ward oil cycle threatens the existing and future readiness of these applications.

2.13 SUMMARY

This research, therefore, takes the role of an investigation of how applied artificial intelligence and machine learning techniques can be adapted in well logging towards solving some of the various challenges that log data gives, especially missing or inconsistent log data. Literature review of historical context of wireline logging is presented, focusing on its importance in characterization of hydrocarbon reservoir. It looks at a range of methodologies from conventional statistical approaches to modern AI techniques that include ANNs and SVMs. Some significant problems encountered in petrophysical analysis are data quality and regular missing log data, which makes the uncertainty in reservoir characterization. Existing log reconstruction techniques are critically analyzed, demonstrating that integration is necessary to create efficient methods for the analysis of data. The review also exhibits a research need area which may include the missing standard techniques of specific applications, the models lack versatility for the application in different geological conditions, and hybrid methodologies are not well explored. It will henceforth develop an Integrated Approach on AI/ML Assisted Log Reconstruction Methodology that will ensure the characterization of accuracy and reliability of a reservoir on shortcomings witnessed in literature. The proposed methodology combining insights derived from this study will improve decision-making processes in the oil and gas industry.

CHAPTER 3
RESEARCH METHODOLOGY

3.1 RECONSTRUCTION OF WELL LOGS USING THE PROPOSED INTEGRATED FRAMEWORK OF MACHINE LEARNING TECHNIQUES

Reconstruction of logs using the proposed framework begins with data acquisition and pre-processing, using again for this purpose the Volve Field dataset located in North Sea. This is a publicly sourced dataset accessed via the Norwegian Petroleum Directorate's DISKOS database and contains crucial well log data, sonic (AC), density (DEN), neutron porosity (NEU), and resistivity logs (RDEP, RMED) and others. The first step is to examine the quality of data, where missing logs, inconsistent data and noise are identified. Because missing data occurs pretty normally with any logging operation due to tool failure or a bad wellbore condition, this step flags incomplete logs for further analysis. This step is also used to identify the relationship between the variables and ensure that the data set is clean and properly formatted. This process then allows one to have adequately prepared data ready for machine learning applications in laying an apt basis for proper and reliable modeling. Outlier detection is a part of the workflow, because it prevents the skewed performance of the model due to anomalous data points. For this purpose, three algorithms are used LOF, One-Class SVM, and IF. As for LOF, it works based upon the analysis of a local density measurement. It identifies outliers in those points that have highly localized density compared with their neighbouring points. One-Class SVM separates the outliers as a class from the majority of the data points. Isolation Forest works on the basis of randomly partitioning the dataset; however, it identifies data points with fewer splits so as to isolate them as outliers. These algorithms work in synchronization to add strength to the dataset by reducing noise and ensuring that the remaining data really speaks about the characteristics within the well log. Having identified the outliers, the correlation analysis is done by applying some of these techniques: correlation coefficient, PCA, and variance threshold plots.

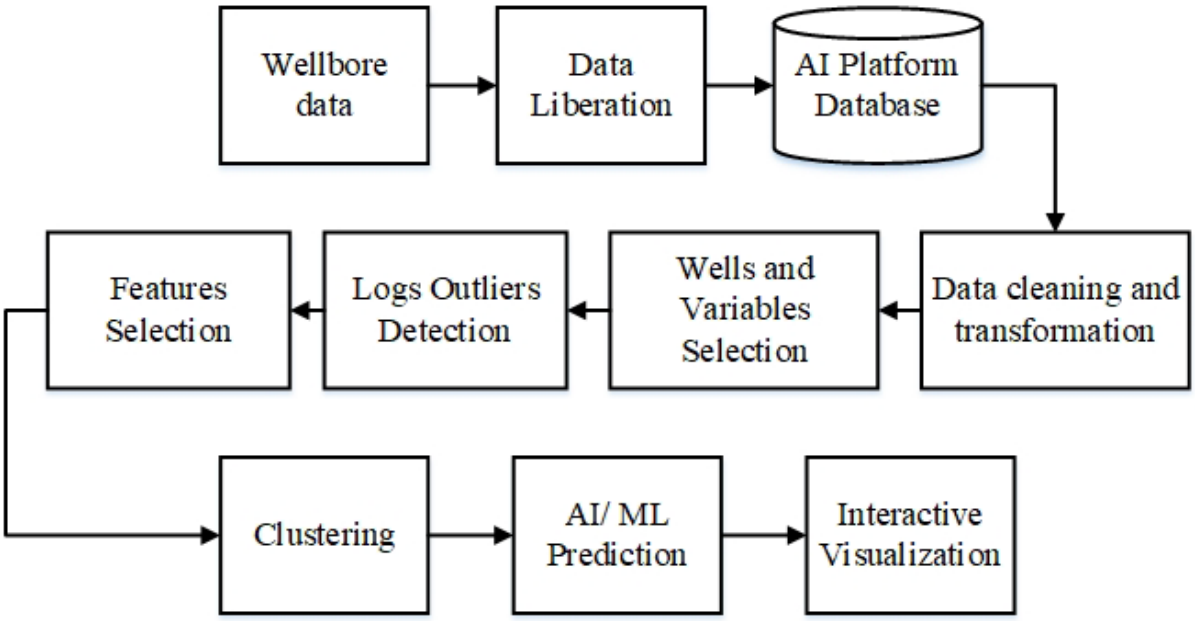


Figure 18: Proposed Framework Block Diagram

The above technique selects the most significant variables relevant to the log reconstruction process. Sonic (AC) logs correlate strongly with DEN, NEU, and PEF; hence it can be useful for cluster and machine learning schemes. Contrary to this, variables such as RMED, RDEP, and GR exhibit moderate low and noisy correlations and can, therefore, be completely omitted from the study in the future. The reduced subset of variables, such as AC, DEN, and NEU, becomes a basis for subsequent clustering and interpretation with the help of machine learning. In that framework, the application of GMM-based clustering is used. GMM is the probability model that decides clusters in an elongated shape and gives data points membership in a cluster by the probability belonging to each cluster. Based on BIC and AIC scores, the framework decides the number of clusters. The GMM method is a more robust approach to the clustering because it assigns every data point to two different clusters based on the top two probabilities and thereby improves the subsequent predictions. After clustering the data, it's used in the Random Forest algorithm to predict missing log data. It is a type of ensemble learning technique that builds multiple regression trees by averaging their output as the final prediction. In its implementation, cross-validation and hyperparameter tuning are used to increase the model's performance. The dataset is split using five-fold cross-validation and the trained model saved for predicting missing or outlier data. For every sample in the prediction dataset, the framework computes the top two clusters uses their ML models to generate two predictions per sample, and then incorporates these as the final prediction. Putting both together, the final prediction is computed as a weighted average of the two predictions with normalized

probabilities as weights. The framework also computes uncertainty with the predictions; thus, further validation to the outcomes is achieved. Using clustering and machine learning techniques this framework offers robust solutions to log reconstruction issues especially when the input data are either missing or noisy.

3.1.1 KEY LOGS FOR MACHINE LEARNING ANALYSIS

The following sonic log (AC), bulk density log (DEN), and neutron porosity log (NEU) are identified as key variables for ML-based clustering and log reconstruction to accomplish the above study. These logs have been selected on rationality, such as having strong direct correlations with the reservoir properties, porosity, and lithology, which have to be clear and accurate for reservoir characterization. Other logs, such as gamma ray (GR) and resistivity logs (RDEP and RSFL), are less correlated with target log(s) and sometimes add noise to the models. They were verified using correlation coefficient analysis, principal component analysis (PCA), and variance threshold plots where lesser relevant logs were filtered out to reduce noise and improve the accuracy of the model.

Another quality control measure was a randomly generated variable which validated the effectiveness of the correlation methods applied. Comparing the correlation of the real logs with a random variable it was established that lower correlation logs-GR, RDEP and RSFL-did not significantly contribute to improving the model and were therefore dropped from analysis. This filtration process is therefore important in ensuring that selected logs, especially AC, DEN, and NEU, give a strong basis for ML-based log reconstruction techniques. The major divisions of the whole AI/ML based well logs QC and reconstruction scheme Wellbore platform where in the form of a database, 2. AI platform where freed data passes through various AI/ML based workflows and, finally 3. User interface (UI) Software that serves primarily as the visualization layer and integrated across the previous two steps.

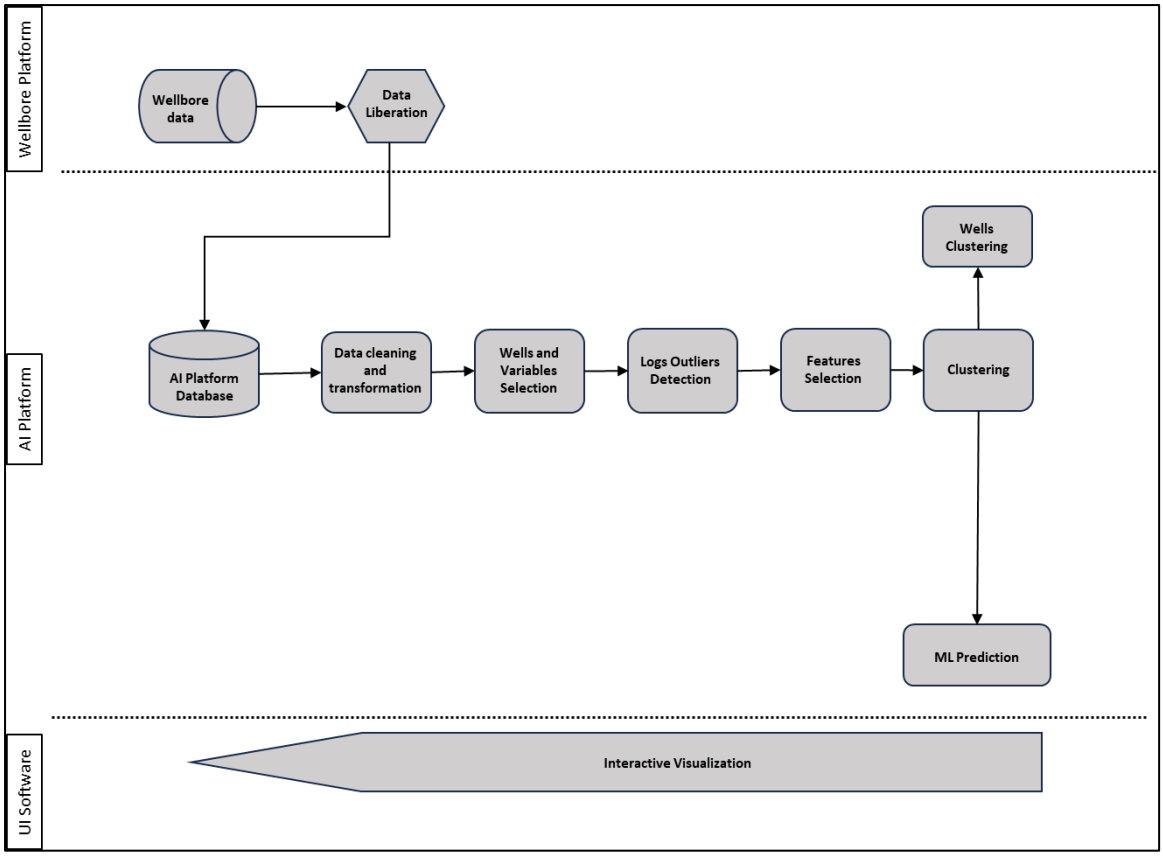


Figure 19: AI/ML Based LogQC, editing and reconstruction workflow.

The figure 19 depicts the structure of AI-based well log reconstruction and the procedure to prepare data for ML-based prediction. It has been structured in a step-wise manner starting from raw data acquisition all the way to the final ML prediction and interactive visualization.

i. Wellbore Data and Data Liberation

The data is first liberated in the form of wellbore data, which is essential well logs acquired during the drilling operation. Well logs are a record of the geophysical measurements captured during the drilling process, which provides information about what formations have been drilled. Among these logs, key data are contained in compressional slowness, or sonic logs; density; neutron porosity; gamma ray logs; and several types of resistivity logs, all of which help describe the subsurface. Data liberation is getting raw data from the well logs out of their storage and making that available to the next operation for subsequent processing. In most cases, well log data is archived in proprietary formats or distributed on scattered databases. Thus, data liberation ensures that the raw data is standardized and transformed into formats that can be easily exploited in further analysis in the AI platform.

ii. AI Platform Database

Once wellbore data is liberated, it is archived in an AI platform database. A centralized repository of all the well logs and relevant data from multiple wells, such a database stores, controls, and provides access to huge data sets of wellbore data in a way that supports easy queries as well as retrieval. AI platform database also supports parallel processing, enabling it to deal with big data, which is important when complex oil fields are involved with many wells. The database also allows for seamless interoperability with a variety of AI and machine learning tools that would be utilized during subsequent steps. It is also important to store the well log data in an organized and structured manner to ensure easier access for cleaning, transformation, and predictive modeling.

iii. Data Cleaning and Transformation

Once data is stored in the AI platform, data cleaning and transformation is the subsequent step. Well log data typically possesses noise, missing values, or inconsistency in the data themselves. This step cleansed raw well logs from the presence of outliers, errors, and missing data points. Techniques of interpolation, smoothing, and normalization applied to make the data consistent in form for further analysis.

Data Transformation

Data transformation is the process of converting data to readable formats for the machine learning algorithm. Such transformations include making variables scale normal, encoding categories, and others. Proper cleaning and transformation ensure that the dataset taken to extract features and model is of high quality.

iv. Wells and Variables Selection

Next to this cleaning is the wells and variables selection. Not all wells or all variables (log types) need to be utilized in the reconstruction process. In this step, among the most pertinent wells and log variables needed for proper log reconstruction as well as those to be used in building predictive models are selected. This is a data-driven process, and usually, statistical methods and/or domain experience is employed when ascertaining wells and variables that contribute significantly to the model. For example, some of the logs such as compressional slowness (AC) could have highly correlation with density logs (DEN) and therefore chosen for reconstruction while others with low relevance are discarded. This step benefits in reducing the time taken to compute and ensuring that the variables picked are highly informative to the model.

v. Detection of Log Outliers

Critical log outlier detection: In this step, all data points appearing anomalous which could distort the behavior of final machine models need to be caught and removed. Well logs are normally heavily populated with outliers due to equipment failures or geological anomalies or errors in collecting the data. Outlier detection ensures that only accurate and reliable data should be left for predictive modeling. LOF, One-Class SVM, and Isolation Forest (IF) are some of the classic outlier detection methods. Most of these algorithms often work based on either querying the local density of points or partitioning techniques to isolate outliers. The model is strong enough and also resistant to bias from outliers when it excludes such anomalous data points during model formation; in turn, this improves the accuracy of predictions following those models.

vi. Features Selection

After the handling of the outliers, feature selection follows. In machine learning, not all features are useful in predictive modeling. Feature selection is a process of selecting the most relevant features that can carry the highest power of prediction for a certain task, and in our case, this is log reconstruction. Such a process reduces dimensionality on a set of datasets, and it also improves the model's behavior with lesser computational complexity. Feature selection techniques, such as PCA, variance thresholding, or correlation analysis, can be used to filter out the most important variables while discarding the redundant or unimportant ones. This phase is important in building a parsimonious model which generalizes well for new data.

vii. Clustering and Wells Clustering

After feature selection, perform clustering to group the wells into categories based on similarities found in their log data. Using methods like GMM (Gaussian Mixture Model) for clustering, each set of wells with similar log characteristics can be clustered into their respective clusters. This would be useful as it will highlight the inter-relations between various wells, but it will apply the same information to all recognized geological patterns or trends within the data. This way, clustering also allows for more focused machine learning models by grouping together similar sets of wells. For example, it is much more likely to obtain similar patterns of missing logs between wells in the same cluster and, therefore, would lead to a better reconstruction. Clustering also helps in identifying populations of wells that show similar trends; this is an important factor in regional geology as well as in the characterization of the reservoir.

viii. ML Prediction

The above-obtained well data will be readied for ML prediction following clustering. Predicting missing well log data and reconstructing incomplete or noisy logs is done using an ML algorithm like Random Forest. Using the cleaned and formed clusters, the model trains on the transformed dataset. Techniques like cross-validation ensure and averages their output to make a prediction, helping avoid overfitting and increasing accuracy. It also employs hyperparameter tuning to optimize the performance of the model. Thus, the final ML prediction comes out to be a reconstructed log dataset that fills in missing data and noisy logs with patterns learnt from the dataset.

ix. Interactive Visualization

The last step in the framework is interactive visualization, where the results of the log reconstruction and ML predictions are presented in easily interpretable and interactive format. Visualization tools enable the user to explore the data and model outputs so that predictions may be validated and relationships between variables known in detail. Interactive dashboards can be produced which simultaneously displays reconstructed well logs side-by-side with original logs, highlight regions where the data are missing or noisy, and enable filtering of data by well or variable. This is important in communicating the outcomes of the log reconstruction process and aiding domain experts in making informed decisions based on the data. Interactive visualization also fosters collaboration among data scientists and geologists, so that the reconstructed logs are credible and robust enough for further geological analysis.

3.2 DATASET DESCRIPTION

Dataset employed in the proposed framework: Volve Field - comprises high-resolution data taken along wellbores from various oil and gas exploration exercises. This data set is publicly available in the DISKOS offered by the Norwegian Petroleum Directorate. The dataset offers critical measurements in well logs- that are inevitable for subsurface characterization and more downstream analysis. In general, the dataset is assembled from several boreholes drilled into what geologically is a very complex environment, comprising many faults and clearly multi-sandstone source formation Middle Jurassic in age. This forms logs of a sonic, density, neutron porosity, and resistivity, which are now assembled to provide the geological properties and conditions of the subsurface.

Reflective conditions during logging operations-the data typically represent missing logs, such as tool failures or adverse wellbore conditions, which requires preprocessing and data cleaning to obtain reliable and accurate results on subsequent analyses. Metadata about drilling operations, such as rate of penetration and calliper measurements, help in gaining insight to the drilling environment and the integrity of the wellbore. This much wealthier dataset forms the basis of proposed framework attempts to apply more advanced techniques from machine learning in log reconstruction, outlier detection, and well clustering that facilitates an enhanced understanding of reservoir characteristics and tailor's hydrocarbon exploration strategies.

3.2.1 ATTRIBUTES OF DATASET

This conceptual framework of data used for reconstruction well log includes well bore data, which is very important in the characterization of subsurface and comprises various geophysical measurements that are taken during drilling. This dataset usually comes from oil and gas exploration projects and covers information from multiple wells. The reconstruction of well log was comprised by the following primary attributes or features found within the dataset:

Depth: This is the measured depth at which the logs were made. In most measurements, the depth will be in meters. It is a reference parameter since all of the log measurements link them to a definite geological formation.

Compressional Slowness (AC-Sonic Log): This log shows the time it requires for the travel of compressional sound wave through the formation; usually, in microseconds per foot. It will be able to describe porosity and lithology in subsurface layers.

Bulk Density (DEN): DEN captures the log of density of the rock formation. The measurement is taken in g/cm³. Interpretation: This is used to infer porosity of the rock. This is critical for reservoir characterization and estimation of hydrocarbons in place.

Gamma Ray (GR): The gamma-ray log measures natural radioactivity of the formation. It records in API units. GR helps differentiate between shale, naturally and normally more radioactive, from a cleaner reservoir like sandstone or limestone. This log is so crucial to recognising lithology and formation boundaries.

Resistivity Logs (RDEP, RMED): Resistivity logs measure the electrical resistivity of the formation and can be recorded in ohm-meters. RDEP is a deep resistivity log that measures far away from the wellbore, whereas RMED measures resistivity nearer to the wellbore. It is

important to identify hydrocarbon-bearing zones as hydrocarbons have higher resistivity than saturated formations by water.

Neutron Porosity NEU: Neutron Porosity measures the hydrogen content in the formation. The response is reported in porosity units or fractions. It offers an estimate of the rock's porosity, more so in water- or hydrocarbon-bearing formations since hydrogen is mostly contained in fluids.

Photoelectric Factor (PEF): This is one of those properties that would give more information related to mineralogy, primarily about the gamma ray absorption of the formation at low energies. The PEF log is very useful in distinguishing between different matrices of rocks-limestone, for example, and sandstone.

Rate of Penetration (ROP): This log represents the speed at which the drill bit is penetrating the rock formation, measured in meters per hour (m/h). The data indirectly speak about the hardness of the formation and can be related to changes in lithology.

Bit Size (BS): It is the diameter of a drill bit used to make a well. It proves to be very useful in the interpretation of well logs, particularly caliper logs, and provides accurate geometry for the wellbore.

Caliper (CAL): The caliper log measures the wellbore diameter and depicts irregularities in borehole size. Therefore, it helps to determine washouts or caving in the wellbore, which can alter the accuracy of other logs.

Survey and Wellpath (SURVEY/TL Well Path): These attributes describe well trajectory and directional drilling, especially in relation to deviated or horizontal wells. They are very important for understanding the spatial orientation of a well and its relationships to subsurface formations.

Zonation (ZONATION_1): It is an attribute that is descriptive of geological zonation or layer information that describes the geological formations or zones encountered at different depths. Understanding the sub surface’s stratigraphy and depositional environments is assisted with the help of zonation.

The attributes provided thus allow having a complete view of the subsurface, thereby giving a detailed geological and reservoir analysis.

Thus, in this sense, the framework pre-processed these logs about noise and outliers as well as missing data before they were applied to the machine learning models that would be applied on missing values, reconstruction of incomplete logs, and cluster wells that could be projected based upon geophysical characteristics. Therefore, the diversity of attributes of the dataset ensures that all the geological and petrophysical properties necessary for the reconstruction of wells' logs and subsurface characterization are included.

3.2.2 DATA VOLVE FIELD DATASET- WELL LOG RECONSTRUCTION

It is a dataset retrieved at the central part of the North Sea at the south extremity of the Norwegian sector, that has been used as a starting point for research and applications of machine learning (ML) techniques towards well log reconstruction. Since the oil produced from this reservoir is thought to be of Middle Jurassic sandstone, specifically Hugin Formation, therefore, it makes a very vital role in the geological significance of hydrocarbon exploration in the region. With a calculated 173 million barrels of oil in place, this field holds sufficient information that could be applied to ML algorithms toward the characterization of the reservoir and predicting it. The field was discovered in 1993, with approval of the development plan in 2005. It reached full production at the onset of 2008, reaching its ultimate oil production rate 56,000 bbl/day. More wells were drilled between 2012 and 2013 to achieve total recovery, increasing the field's operational life. Volve field offers well logs, production data, and long-term records; hence, this makes it a highly suitable field for analysis with ML-based approach that goes into outlier detection, followed by clustering, logs reconstruction, etc.

3.2.3 DATASET OVERVIEW AND AVAILABILITY

The data of the Volve field is available on Norway's DISKOS National Data Repository. This dataset consists of complete well logs and core data, as well as production records along with other various geological and geophysical datasets. Because the dataset from the Norwegian Petroleum Directorate is public and accessible, it provides a great opportunity for researchers to develop data-driven models based on such richness in data. With this richness in data, the researchers can go into almost every aspect of reservoir performance by the help of advanced ML techniques for predicting key parameters of reservoirs. These are logs from several wells, and some of the measurements they contain are sonic (AC), bulk density (DEN), neutron porosity (NEU), gamma ray (GR), resistivity-both deep resistivity (RDEP) and microspherical focused resistivity (RSFL) and photoelectric effect (PEF). All of these logs are essential to infer the various petrophysical properties in subsurface formations such as porosity, permeability, and fluid saturation.

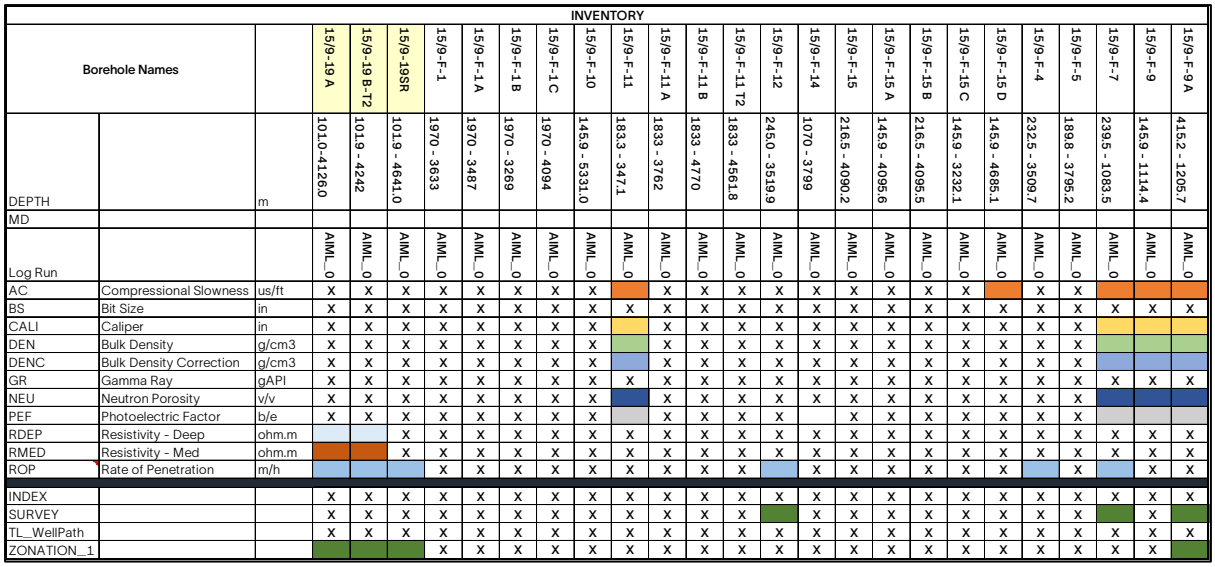


Figure 20: Data outline- 4TB of data, Well Logs, Seismic, Production

Figure 20 provides summary of inventory of availability of well logs for different boreholes. Top rows naming the boreholes are labeled as "15/9-F-1 SR", "15/9-F-1A", etc which are the borehole names. The column on the left provides type of log information: "AC" (Compressional Slowness), "BS" (Bit Size), "CAL" (Caliper), "DEN" (Bulk Density), "GR" (Gamma Ray), "NEU" (Neutron Porosity), "RDEP" (Resistivity - Deep) and "RMED" (Resistivity - Medium). These logs measure various physical properties of the boreholes, such as resistivity, density, and porosity, that are vital to the well log analysis. The structure of the matrix depicts whether every type of log is available for a different well, denoted through coloured cells. For instance,

- Orange-coloured cells appear to depict the AC (Compressional Slowness) logs available for particular wells, including wells 15/9-F-5, 15/9-F-7, and 15/9-F-9A.
- Yellow and blue cells make use of density-logs DEN and DENC, meaning that the logs apply to many wells, including 15/9-F-11, 15/9-F-12, and 15/9-F-5.
- Gray cells signify the log for Resistivity, which may involve RDEP and RMED, meaning that not all wells have the same.
- The green cells refer to Zonation_1 and TL_WellPath, which will prove very useful in correlating zones in geology or simply mapping the trajectory of the well.
- The blank cells may indicate that some log data is missing for certain wells, hence this figure captures where the log data could be incomplete. This figure is likely an inventory that keeps track of which well logs are available for which boreholes, thus aiding the planning of further analysis, preprocessing, and reconstruction of data work.

3.2.4 WELLBORE PLATFORM

Wellbore Platform: This is the core of the system architecture and data will be loaded into an industry-standard wellbore interpretation platform, and there should be available and harmonized basic petrophysical parameters, such as well name, borehole name, depth information, well log curve name. A total of 24 borehole data were uploaded with 11 log curves per well, including zone and deviation survey information, as summarized in Table 1. To ascertain that the database remains harmonized between wells, the interpreted data from wellbore is released to the AI platform for design, deployment, and management of AI/ML logic and model.

3.2.4.1 AI Platform

The AI Platform database comprises a highly curated collection of wellbore data, thus basically very vital for proper and efficient modeling.

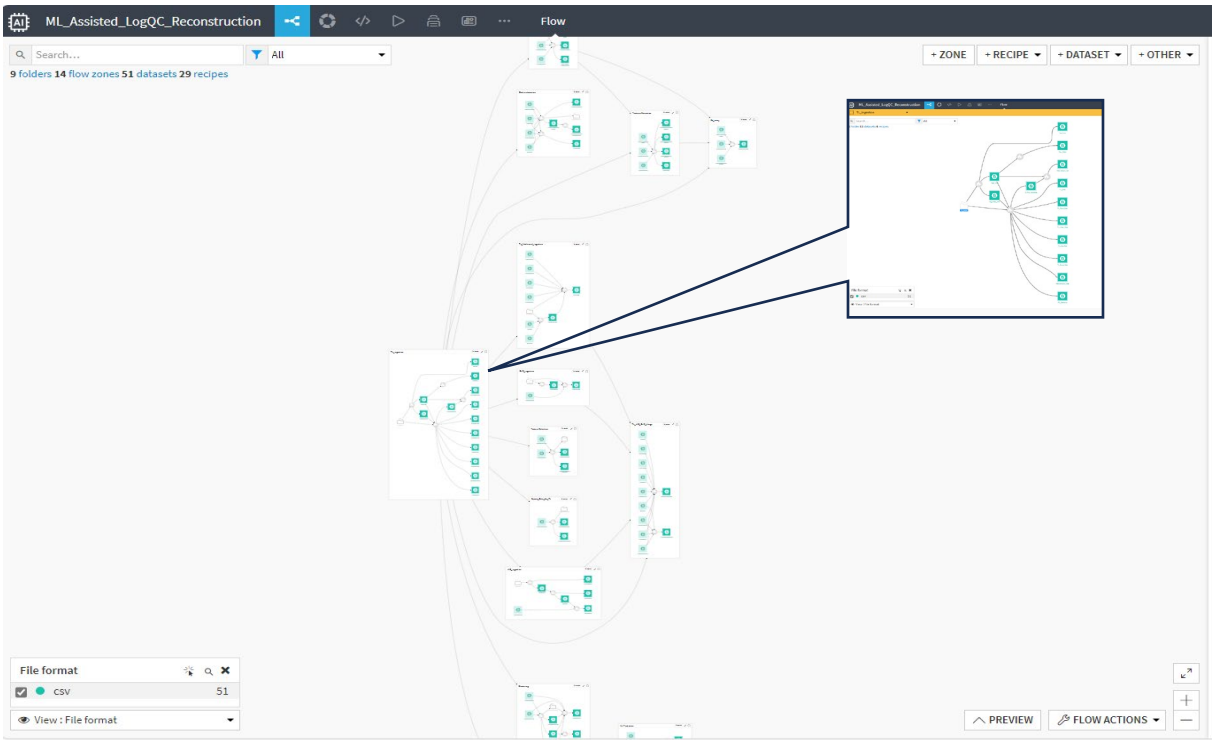


Figure 21: Underlying AI platform workflow

The database contains seven important log curves, particularly selected around issues that have significant importance in succeeding pre-processing and processing stages of analysis. These log curves include Bulk Density (RHOB), used to evaluate physical properties of the rock; Compressional Slowness (AC) helps in understanding wave propagation through geological formations; Deep Resistivity (RDEP), used for identifying hydrocarbon zones; Gamma Ray

(GR), which delivers valuable insights into the lithology of the formations; Medium Resistivity (RMED), complimenting fluid content analyses; Neutron Porosity (NEU), which is essential in estimating porosity; and Photoelectric Factor (PEF), which enhances mineralogical interpretation. Through this wealth of variety in log data, AI Platform Database is empowered to support intricate machine learning techniques which enable proper data pre-processing and anomalous data detection with thorough analysis.

a. AI Platform Database

In general, the overall approach to this integrated scheme leads to superior accuracy of reconstructed logs and greater characterization of subsurface that is the sound hydrocarbon exploration as well as production strategy. With multiple types of data integration and the application of advanced algorithms, this platform would be able to be applied to complex geological analysis, and through proximal learning, would assist with informed decision making.

3.3 DATA PRE-PROCESSING

Pre-processing is those important steps preparatory for the analytic of borehole logging data, especially when applying it in the use of artificial intelligence and machine learning. Generally, the data pre-processing workflow involves data cleaning, transformation, and liberation. Each one of them undergoes such a way that they minimize risks of errors, therefore leading to quality increases from the dataset analyzed. Thus, the pre-processing steps involved in data are as follows

- **Data Cleaning and Transformation**
- **Data Clipping**

Data pre-processing is essential in preparing borehole logging data in the analysis process, especially when AI and machine learning algorithms are being used.

Data cleaning and transformation mean that it will have washed out any anomalies that may distort the result. It includes removing log values that exist outside the expected physical limits. Examples include a negative value for resistivity or induction, which are physically impossible measures. Logarithmic transformations are also applied to variables like resistivity and porosity so as to make variance stabilized and approximately normally distributed; this favours the performance of machine learning algorithms. Statistical methods are used to detect outliers, thus eliminating such erroneous or irrelevant data points either by adjustment or exclusion. In

addition, constant values over extensive depths are eliminated for adding no analytical value. Data clipping is the final step of pre-processing and it forces a limit on log values so that they fall within some predefined range.

3.3.1 DATA CLEANING AND TRANSFORMATION

Data cleaning and transformation are the first processing steps along the pipeline to ensure the dataset doesn't hold any anomalies that may have skewed the results obtained from the pipeline. After all transformation, the first process toward data clipping is to identify any log value that occurs outside of the expected physical limits and eliminate it. For example, in such logarithmic transformation usually applied to the dataset further more refines it, resistivity values cannot be negative as it computes up to which extent a material resists the flow of electric current, value that cannot exist as negative. Such wrong values will produce drastic errors in the system when feeded to statistical or ML tools. This transformation is very useful in case of those variables whose behavior can be described by an exponential growth or, more generally, has a heavily skewed distribution like resistivity and porosity. The log of such values improves stability of variance brings data closer to normal distribution. This will make the machine learning algorithms work better because most of the machine learning algorithms tend to need normal distribution data for optimum performance.

This transformation also compresses the range of extreme values, which reduces outlier influence for easier modeling of data relations. Further data cleaning issues are the identification of obvious outliers-values that fall significantly outside of the bulk of data points. Data cleaning and transformation are essential parts of the processing pipeline of borehole logging data so that the dataset does not contain anomalies which could skew analytical results. The first step is to identify and eliminate log values that are beyond the limits as physically expected. For instance, resistivity values, which are indicative of resistance to the flow of electric current, ought to be nonnegative. Negative resistivity values bring no physical meaning and might generate large errors if the data processing is carried out using statistical or ML-based tools. In order to enhance the quality of data, logarithmic transformations are generally applied, especially for such properties with exponential growth or very heavy tailed distributions, for example, resistivity and porosity. Below is the mathematical representation of the logarithmic transformation represented in the eqn (1):

$$Transformed\ Value = log(x) \tag{1}$$

Where x is the original log value. This transformation stabilizes variance and shifts the data closer to a normal distribution, which is crucial as many machine learning algorithms perform optimally with normally distributed data. Additionally, it compresses the range of extreme values, thereby reducing the influence of outliers and facilitating easier modeling of data relationships. Where x is the original log value.

This stabilization of variance pushes the data toward the more normal distribution. This is important because many machine learning algorithms expect, and even perform best with, normally distributed data. It also compresses the range of extreme values, thereby reducing the influence of outliers and making easier modeling of the data relationships. This process typically involves translation of proprietary formats native to logging software into standardized data formats, which could be much easier to exploit using AI/ML frameworks. Then, the released data should be ordered in an appropriate manner whereby all measurements of the logs are uniformly spread as far as depth and resistivity measurements and porosity are concerned. This would then make its integration into the ML pipeline easy, which would subsequently make all activities regarding modeling and analysis much more efficient. The other aspect of data liberation is assurance against data loss or corruption as it is transferred. Integrity checks of the data are done which, in essence means verifying that the released set of data reflects the original data pertinent to each measurement with its metadata. Maintaining the integrity of the data ensures one can place trust in the outcomes generated by AI/ML applications. It is only when complete data cleaning and transformation are added to the productive data liberation that the bases will be sturdy enough to help bear the application of AI and ML techniques in the analysis of borehole logging data. This goes beyond enhancing quality to be provided with available data but also makes easy the process of transition from raw data towards actionable insight, supporting better characterization and decision-making processes within the oil and gas industry.

b) Data Cleansing and Transformation.

In this stage, some basic mathematical formulae were designed that enabled instant QC of the data consisting of clipping of data (when data at a given location exceeds the provided range), identification and removal of anomalous log values, and removal of constant continuous values or logarithmic transformation where certain workflows in predictions like permeability could not be done. This step could apply data QC to all of the wells listed below in under a second, so would be a huge efficiency booster for the interpreter. Anomalous log values must be flagged

so that data integrity is assured. The most common way an anomaly is detected is through the use of the z-score, which is the number of standard deviations a data point is from the mean. The equation for the z-score is:

$$Z = \frac{x - \mu}{\sigma} \tag{2}$$

Where:

z is the z-score of the log value.

x is the log value being evaluated.

μ is the mean of the log values.

σ is the standard deviation of the log values.

Typically, a threshold (commonly (z > 3) or (z < -3)) is set to classify values as outliers. If the z-score of a log value exceeds this threshold, it is marked for removal from the dataset.

3.3.2 REMOVAL OF CONSTANT CONTINUOUS VALUES

Another way of eliminating constant continuous values-the observations that are the same across a large depth interval-the method monitors the variance of each log parameter. A flag of elimination can be assigned with a zero or less than a set tolerance level. The variance V for a log parameter x_i may be calculated using the eqn (3):

$$V = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2 \tag{3}$$

Where:

N is the number of observations.

$\bar{x}$ is the mean of the log parameter values.

If V < Threshold, the parameter is removed from the dataset.

Logarithmic Transformation

Where in some prediction processes needs to be carried out in a logarithmic form, for example when estimating permiability, it is transformed by the eqn (4):

$$y = log_b(x) \tag{4}$$

Where:

y is the transformed value.

x is the original log value.

b is the base of the logarithm (commonly 10 or e).

This transformation has the effect of stabilizing the variance and making the distribution of skewed data more normal. This is particularly helpful when dealing with features like permeability, which tend to exhibit exponential characteristics.

Efficiency Gains

These operations are actually carried out mathematically QC automatically by the process, and it evaluates all of the listed wells in less than a second. This would mean an important improvement in efficiency, enabling better time usage on analysis and less on data validation by the interpreters, thus accelerating the decision-making process in the characterization and management of reservoirs. With automation and this well-structured approach, not only is data reliability enhanced, but even its proper application in subsequent analyses becomes clear through models involving machine learning. During the data pre-processing stage, crucial mathematical equations and algorithms are utilized for the proper execution of rapid quality checking on the borehole logging data; hence, it enhances the efficiency as well as the reliability of the dataset that is to be applied for further analysis or machine learning applications. These fundamental operations include clipping, anomaly detection, removal of constant values, and logarithmic transformation.

3.3.3 DATA CLIPPING

Data clipping is a pre-processing technique that calls for all measurements in one dataset to fall within acceptable limits that are predefined. In the scenario of borehole logging data analysis, it does play an important role in ensuring the integrity and reliability of the dataset before going for further analysis or application in machine learning. Data Clipping: A process of finding a Lower Limit and Upper Limit that specifies the value range for data point occurrence represented by the eqn (5):

$$x' = \begin{cases} Lower\ Limit & if\ x < Lower\ Limit \\ x & if\ Lower\ Limit \leq x \leq UpperLimit \\ UpperLimit & if\ x > Upper\ Limit \end{cases} \quad (5)$$

Where:

x is the original log value.

x' is the clipped log value.

$Lower\ Limit$ and $Upper\ Limit$ define the acceptable range of measurements.

Lower Limit and Upper Limit identify the limit measurement accepted.

This formula ensures that any log value dropping below the Lower Limit gets clipped to the Lower Limit, whereas values above the upper limit get clipped at this higher limit. It also helps to ensure that any log values which fall below the Lower Limit are raised up to it while those exceeding the Upper Limit will be capped at this maximum limit. This, therefore, effectively reduces extreme outlier values that might provide skewed results, hence rendering a much cleaner and more focused dataset. Data clipping is especially helpful for such applications as borehole logging where some physical constraints predict the bounds of certain expected values. For instance, the resistivity measurements, which denote the subsurface material's electrical conductivity, should be confined to some degree; otherwise, a value might be an indicator of measurement error or abnormality. Clipping the data in that sense will make sure that the dataset is not just physically meaningful but also manageable statistically. Furthermore, data clipping also improves performance of most statistical analysis and machine learning algorithms, in that such algorithms are mainly designed based on the assumption that data is normally distributed. Clipping reduces the impact of outliers so that nearly normal distributions of data can be achieved, thereby improving subsequent analytical robustness. In short, data clipping is one of the critical steps in data preprocessing, which not only improves the quality and the reliability of the dataset but also supports the conduct of more accurate and efficient analysis in the proposed framework.

3.4 DETECTING OUTLIER USING THREE ML ALGORITHMS

Outlier detection is an important sub-process within the data pre-processing phase of the proposed framework as this helps identify and correct anomalous values on well log data. The existence of outliers affects the performance of machine learning (ML) models, especially for predictive tasks in reservoir property estimation. The source may emanate from measurement errors, data entry mistakes, or other subsurface conditions that are unusual, hence values that are significantly different from the general distribution of the dataset (Wu, Jain, Kulkarni, & Abubakar, 2018). Therefore, the detection of these outliers is crucial because it ensures that a dataset is well representative of the underlying geology phenomena (Wang, Bah, & Hammad,

2019) and goes well with reliable training and predictions. For outlier detection, the following three algorithms were considered which are;

- **One-Class Support Vector Machine**
- **Isolation Forests**
- **Local Outlier Factor**

This is an approach that focuses more on identifying the boundary of normal data points through the construction of a hyperplane that separates most data points from outliers (Talreja, 2023). Mapping the data into a high-dimensional feature space will enable the algorithm to establish complex structures within the dataset. Such a method is particularly efficient wherever most of the data points are representatives of true distributions, while these outliers are representative of rare occurrences that fall on or outside this boundary (Stuov, et al., 2022). The model learns on normal data to create an optimal decision boundary minimizing the classification error, hence efficiently isolating outliers from the norm.

Isolation Forests (IF): Isolation forest is another strong algorithm applied for outlier detection, mainly in high-dimensional datasets (Singh M. , 2020). It works on the assumption that outliers are more isolated than the normal observations in a dataset. The algorithm creates an ensemble of decision trees, which make random choices about which features to consider and which splitting values to apply in creating partitions within the dataset. The key property of Isolation Forests is that the outliers have fewer splits to isolate than normal points, and hence the path length would be shorter in the decision trees (Misra, Osogba, & Powers, 2019). The algorithm works by computing an anomaly score based on the average path length it takes for isolation of every data point, which makes it flag points of unusual short path length as outliers. This method is efficient, scalable, and capable of handling big data while ensuring to deliver an accurate outlier detection (Jha, Khanal, Seikh, & Lee, 2022).

LOF: Local Outlier Factor: This is a complex algorithm which computes the local density of data points to identify outliers based on their relative proximity to neighboring points. Unlike universal methods, LOF analyzes local data properties that can discover outliers which are not viewable in the overall distribution (Datta, Singh, Singh, Jenamani, & Routray, 2023). The algorithm returns a score for each point as a ratio of the densities of that data point relative to its neighbors; points that have been found with relatively much lower density than their neighbors are raised as outliers. This method is particularly effective in datasets with varying densities as it accounts for local structure of data and not just in global measures. LOF identifies

points that exhibit substantially different local density characteristics and enhances the robustness of outlier detection in complex well log data. In a nutshell, therefore, these three algorithms: One-Class SVM, Isolation Forests, and Local Outlier Factor are thus embedded in the outlier detection process, thereby providing a complete framework in anomaly detection within well log data. Each algorithm gives its own independent contribution in providing a deep exploratory look at the data, and the integrity of the dataset is guaranteed before its use for further detailed analysis and machine learning applications.

3.4.1 LOCAL OUTLIER FACTOR

The Local Outlier Factor algorithm is complex because it essentially presents a density-based outlier detection scheme of the data and works well even if data is not uniformly dense (Banas, McDonald, & Perkins, 2021). While many global outlier detection methods have been devised to measure points in terms of their absolute values or distance from some notion of central tendency, LOF measures each point in space relative to the density of its local neighbourhood. LOF is the core idea of how isolated a point within data is relative to its neighbors when the local density of that particular point is compared with the local density of its neighbors. For a data point x, it can be given by the eqn (6):

$$LOF(X) = \frac{\sum_{y \in N_{k(x)}} \frac{lrd_k(y)}{lrd_k(x)}}{|N_{k(x)}|} \tag{6}$$

Where:

$N_{k(x)}$ represents the set of k-nearest neighbors of x.

$lrd_k(x)$denotes the local reachability density of x.

$lrd_k(y)$ | is the number of neighbors.

The local reachability density is defined as the reciprocal of the average distance from x to its neighbors, thereby quantifying the density of points surrounding x. By evaluating the local density of x in relation to that of its neighbors y, the LOF score provides a normalized measure of how much more isolated x is compared to its peers. When applied in the context of well log data, LOF excels at identifying anomalous measurements, such as unexpectedly high or low resistivity or density values at specific depths. In practice, LOF shows superb capability when there are anomalous log measurements, like significantly high or low resistivity readings or density readings at some depths (Alimohammadi & Chen, 2022). This becomes really

important because when you try to apply geophysical principles, the sub-surface environment could be very heterogeneously different for various formations.

One of the advantages of the LOF algorithm is that it finds subtle outliers in data clusters that are of different densities, thereby potentially picking out anomalies that some algorithms may have missed. Examples include the resistivity data fluctuations caused by complex geological features because it picks out measurements which are significantly different from what is expected for the density pattern, thus potentially picking out errors or significant geological features (Akkurt, et al., An unsupervised machine-learning workflow for outlier detection and log editing with prediction uncertainty, 2023). That would allow LOF to become a good adaptable tool for analysis of well logs, which often contain a wide range of geological conditions and uncertainties in measurements. The major advantage in focusing the method at the local relationships between the points rather than their global characteristics is the robustness it provides to the detection in complicated cases. This capability improves the quality of the data that is used for subsequent applications of machine learning and also contributes to more accurate estimation of reservoir properties and geological interpretation, all constituting a firm foundation to support better-informed decisions to be made in the upstream domain of oil and gas exploration and production.

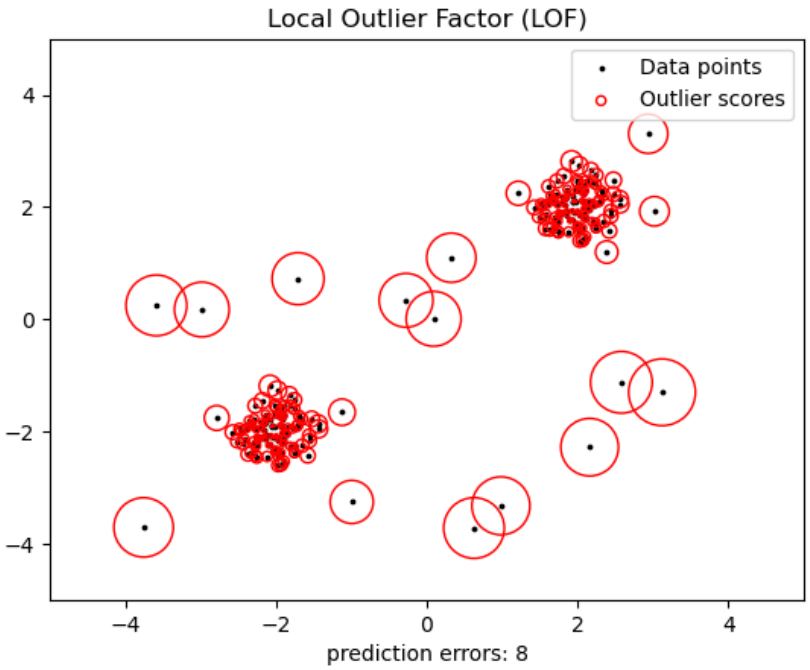


Figure 22: Predicting outliers using LOF Algorithm

- **Data Points**: The red circles are the data points in the data set. It is the observations that LOF analyses in order to find out outliers as shown in figure 22.
- **Outlier Scores**: The black dots are the outlier scores returned by the LOF algorithm. The scores report how much of a point deviates from its local density of neighbors.
- **k-Nearest Neighbors**: The larger circles surrounding some of the data points likely represent the k-nearest neighbors employed in calculating the LOF score. Neighbors are crucial to the determination of the local density surrounding each point.
- **Local Reachability Density (LRD):** For every data point, the local reachability density is calculated by using the distances to the k-nearest neighbors. This density relays information about how lonely or otherwise isolated a point is relative to its neighbors.
- **LOF Calculation**: LOF calculates the score for every single point by taking the comparison of the local density of the given point with the densities of its neighbors. When the local density is greater, then the score is greater for the assigned point; that would mean the point is relatively isolated and more likely to be an outlier.
- **Prediction Errors:** At the bottom left it is marked "prediction errors: 8". This means that the LOF algorithm treats eight points as outliers in this data set.

In short, the above image depicts how LOF Algorithm detects outliers based on the computation of local density of a point with respect to its neighbors. It is more useful in datasets with varying densities since the algorithm can detect slight anomalies that global outlier detection algorithms would miss.

3.4.2 ONE-CLASS SUPPORT VECTOR MACHINE

The proposed outlier detection framework uses the technique of One-Class Support Vector Machine SVM. One-Class SVM is known to be better performing in high dimensional spaces as well; it is also true that complex patterns are more likely to appear in well log data (Akkurt, et al., 2018) (Landauer, Onder, Skopik, & Wurzenberger, 2023) (Farzad & Gulliver, 2020). One-Class SVM is, essentially, a design algorithm. It is meant to differentiate normal data points from outliers by learning the data distribution and determining which deviate significantly from that said distribution. This method is particularly relevant in reservoir characterization studies, where anomaly detection like resistivity or sonic logs abnormality proves to be of great importance in the enhancement of subsurface prediction and further helps in taking much better decisions.

The One-Class SVM is based on the idea of creating a hyperplane in a high dimensional feature space to separate most of the data points (normal ones), from those that are believed to be outliers. It essentially adjusts the boundary around the data points such that most of the points lie within this boundary, or, in other words, the decision function. Those points lying outside the boundary are classified as outliers. Mathematically, the One-Class SVM is an attempt to minimize an objective function that balances the fitting of the model to most of the data points while penalizing the loss of any points outside the decision boundary as it was given by the eqn (7):

$$\min \left(\frac{1}{2}\left|\left|\omega\right|\right|^2 + \frac{1}{vn}\sum_{i=1}^{n} \max\left(0, 1 - (\omega.\emptyset(x_i) - \rho)\right)\right) \quad (7)$$

Where:

$\left|\left|\omega\right|\right|^2$ represents the weight vector that defines the hyperplane.

v is a parameter that controls the trade-off between the fraction of outliers and the margin of the hyperplane.

x_i represents the individual data points.

$\emptyset(x_i)$ is the mapping of the input data to a higher-dimensional space through a kernel function.

ρ is the offset from the origin, controlling the placement of the hyperplane.

In the One-Class SVM, a kernel function \phi(x_i) is utilized to map out the original data into a higher-dimensional feature space so the algorithm can capture the patterns in the data as well as it may reveal non-linear patterns. In detecting anomalies using algorithms, where instances exist in complex data structures, such as in well log data where many variables such as resistivity, gamma ray, and porosity also exhibit non-linear relationships, then the kernel trick may become invaluable (Zhang, Meratnia, & Havinga, 2010) (Yehia, Khattab, Tantawy, & Mahgoub, 2022) (Wurzenberger, Skopik, Fiedler, & Kastner, 2017). In the above framework, the One-Class SVM is trained on a dataset of well log measurements that represent the "good" data points-those that follow expected geological trends. Most of the resistivity, porosity, and sonic log values should show predictable trends based on the subsurface formations they are measuring. This pattern is then learned by one-class SVM, which constructs a decision boundary that encapsulates the well log data in the normal one. After completing the learning step, the model is used to detect the outliers in the dataset. Outliers may be either points located

within a region of the resistivity, where it highly deviates compared to most well log data, or in points where a sonic log has values highly deviating from regular values. The most significant advantage of using One-Class SVM in this context, however, is that it can operate in high-dimensional feature spaces. This is so if we are to take into account several well log variables at once.

With regard to multiple log variables-considering, for example, density, resistivity, and porosity-the algorithm would be much more effective at picking up small anomalies that might not be found when analysing each log in isolation (Sreevidya, 2014) (Suri, M, & Athithan, 2019) (Siwach, 2022). At the practice level, it would mean that data points displaying unusual conditions of subsurface or measurement noises-for instance, substantially high resistivity values that might indicate hydrocarbon presence or values extremely low because of measuring noise-are identified as anomalies by the One-Class SVM. The reasons why this identification is important are as follows: The cleaner the high-quality data, the deeper the task-like additional usage of the model for reservoir property prediction. Outlier detection within the proposed framework is provided in strong form by One-Class SVM, given that it performs well in high-dimensional spaces. The hyperplane-building approach segregating the normal data from outliers helps to identify even the most subtle anomalies in complex geophysical data; hence well log analysis is reliable and quite accurate (Kuna, García-Martinez, & Villatoro, 2014) (Kumar, Khare, Choudhury, & Devaraju, 2024). The use of the kernel functions enables the One Class SVM to model even complex non-linear relationships in the data, and therefore this algorithm is very suitable to be used when detecting outliers in diverse high-dimensional datasets as those often involved in reservoir engineering and geophysical exploration.

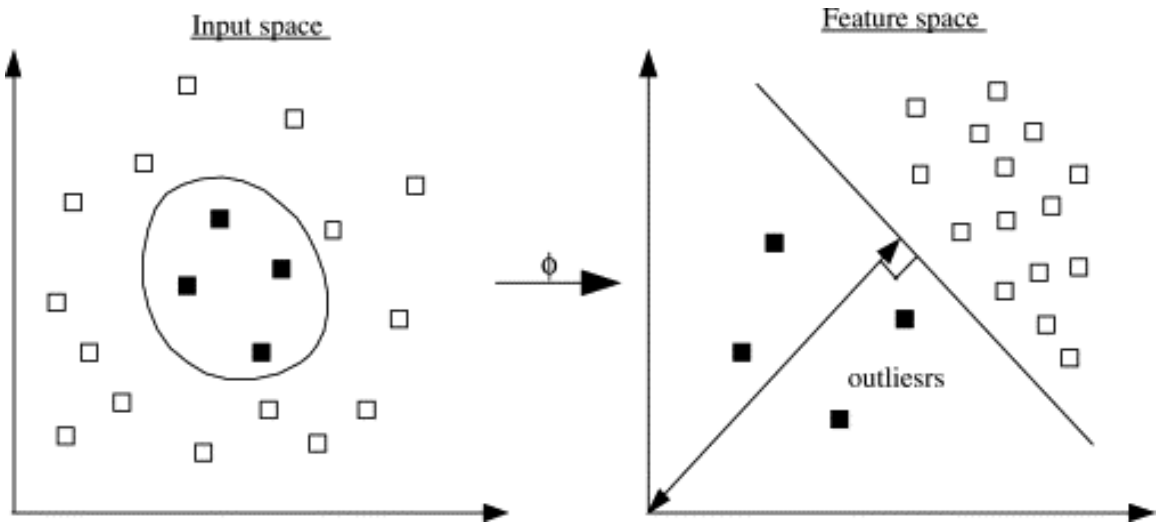


Figure 23: Detecting Outliers using One Class SVM

Input Space: In the framework of the One-Class Support Vector Machine, data first undergoes an analysis in its input space where normal data points and potential outliers can be distinguished graphically as shown by figure 23. The normal data points are represented in figure 23 by squares inside a circular boundary. They are dispersed spatially just like the general pattern of the distribution of the dataset, almost behaving like the data is expected to.

These are then matched to the log measurements of a well and other geophysical information, such as resistivity and sonic log values. The sqaure marked outside the circle are outliers, which are found to be considerably out of the normal distribution. It often becomes very challenging in the input space to differentiate between normal data and outliers because patterns overlap. There also exists in most cases an interdependence among variables through non-linear relationships. Feature Space: In order to accomplish this, one-class SVM maps data from the input space to the feature space with the assistance of a kernel function.

Mapping transformation allows the algorithm to take in the subtlest pattern that appeared much imperceptible in the input space. In the feature space which is of a much higher dimension, a hyper-plane is constructed in such a way that it separates the outliers from normal data points. Thus, I want the placement of the hyperplane in such a way that most of the normal data points lie on one side and the outliers are isolated from each other on the other side. The transformation gets the normal data points packed closer together, thus making it easier to have a clean separation by the hyperplane. Outliers thus, are easier to be detected as they lie on the other side of the hyperplane in this space. Decision Boundary: The decision boundary is, however a critical input and feature space. In the input space, the decision boundary resembles a circular area which encloses most of the normal data points. If the points falling outside this circle are marked as potential outliers. In the feature space, this boundary is represented by the hyperplane. Now, based on this boundary, One-Class SVM effectively separates outliers with normal data.

After the system is trained on normal data, which would correspond to expected geological conditions-for example, typical resistivity or porosity logs-the One-Class SVM is capable of outlining outliers, which could be related to measurement errors or unusual geological conditions or something else in the data. This method translates data into a higher-dimensional space and separates outliers from normal data by a decision boundary. So, it is a robust tool in handling complex, high dimensional datasets in which only clean, reliable data is used for further processing.

3.4.3 ISOLATION FOREST

The IF algorithm represents an efficient, ensemble-based outlier detector and is especially appropriate for large and high-dimensional datasets, as well log data in geophysical applications. While methods such as one-class SVM or Autoencoder to model normal distribution of data IF aims directly to isolate individual data points. This is based on the presumption that anomalous or outliers are of lesser quantity and also distinct from the rest data, hence easier to isolate. The algorithm produces multiple isolation trees, iTrees, which is more or less just a binary tree where the Isolation of every data point is done recursively by splitting the entire dataset. In each split, one feature is randomly chosen and some random value in the range of that feature is chosen as a threshold value to split up that data set. This procedure is repeated recursively until every data point is totally isolated in the (Hodge & Austin, 2004) (He, Zhu, He, & Lyu, 2016).

The basic idea here is that the outliers are very infrequent and dissimilar to most of the data; therefore, they are normally separated much earlier than the normal data points. This is because outliers tend to possess extreme values, which are easier to get separated early in the recursive splitting process. As a result, outliers should, in general, have shorter path lengths in the isolation trees while normal data points will require longer splits to become completely isolated (Grover, 2018). In exchange, the algorithm produces an anomaly score for each data point, showing the efficiency of isolation for the point. The formula for the anomaly score is given by the eqn (8):

$$s(x,n) = 2^{\frac{-E(h(x))}{c(n)}} \tag{8}$$

Where:

$h(x)$ is the path length of point x in the isolation tree.

$E(h(x))$ is the expected path length.

$c(n)$ is the average path length for data of size n the dataset, computed as in the eqn (9):

$$c(n) = 2H(n-1) - \frac{2(n-1)}{n} \tag{9}$$

In this formula, H(n) is the harmonic number that approximates the mean depth of a random binary search tree. score s(x,n) such that it lies between 0 and 1. A score close to 1 indicates that the point is probably an outlier. A data point that has a short path length is likely to have a

high score of anomaly, meaning it got isolated quickly and most probably is an outlier. This comes in handy when working with well log data, for instance, where measurements of resistivity, porosity, and sonic velocity might drastically change along the depth or even among regions. These measurements can include anomalies that either arise from errors in measurement, data entry errors, or special geological conditions such as fractures or non-uniform subsurface shape. For instance, when the resistivity recorded at a depth differs from neighbouring points in a higher or lower value, then either it is because of error in measurements or geological peculiarity (Ghionna, Greco, Guzzo, & Pontieri, 2008). Isolation Forest is very suitable for this type of data since it can provide an efficient solution in handling high-dimensional datasets. Well logs may contain a huge number of features, like resistivity, porosity, and even more geological properties. So, the dimensionality is critical.

IF is known to be very efficient in the processing of large amounts of data; it constructs iTrees which have to perform fewer computational steps than traditional distance- or density-based algorithms. Another remarkable advantage of the Isolation Forest method is its scalability (Sani, Zelst, & Aalst, 2018). Due to the fact that each isolation tree is built using random splits, this algorithm is computationally light and parallelizable, which makes it very efficient for massive datasets. This makes it very suitable for processing large volumes of data in the oil and gas industry, where data from well logs obtained from multiple wells can be enormous. In the new framework, a primary outlier detection technique will isolate abnormal well log data points through Isolation Forest. The algorithm is computationally efficient, capable of handling outliers in complex, high-dimensional datasets.

For example, this might be utilized in detecting the outliers in logs that might have abnormal readings such as unexpected values of resistivity or anomalies in porosity measurements that might distort the performance of the machine learning models. In that regard, the developed framework integrates Isolation Forest with other outlier detection methods, namely Local Outlier Factor (LOF) and One-Class Support Vector Machine (SVM), toward achieving a more robust and comprehensive approach toward the detection and handling of anomalies in well log data. Outliers identified through various methods are corrected or removed to yield the clean, accurate final data set. This finally makes the overall reliability of the models developed for applications such as reservoir characterization and well log reconstruction.

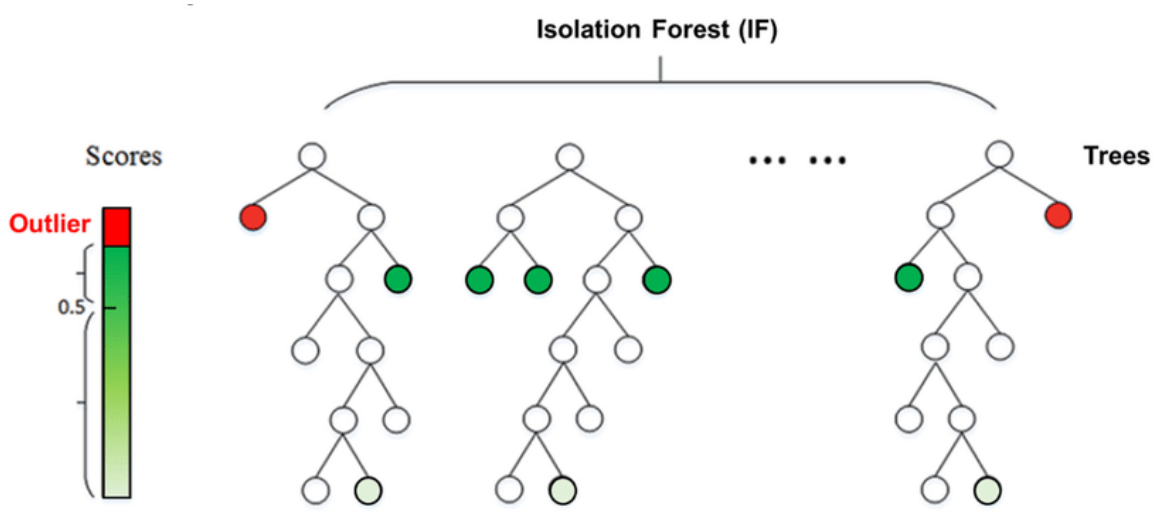


Figure 24: Outlier Detection Using Isolation Forest

The figure 24 provides a visual representation of how the Isolation Forest (IF) algorithm functions outlier detection processes with particular regard to the provided framework. The break up of the elements of the figure into one-by-one rendering their meaning on how the Isolation Forest works.

Trees in Isolation Forest: The Isolation Forest algorithm performs by making an ensemble of two-class binary trees. Each tree is built by recursively splitting the data points with random features and with random split values. In the figure, each vertical structure describes an individual tree where the data points are split at every node. This is because outliers are more easily isolated from the rest of the data, so they would have relatively shorter path lengths inside the tree, or in other words, they are isolated early in the splitting process. The trees in the figure illustrate how different data points move through the nodes of the tree. The branches of the tree split the data points into smaller and more homogeneous groups based on randomly selected features and split points. Path Length and Scoring The path length is the number of splits or steps to separate a data point. Since outliers are unusual or few in numbers, it takes fewer steps to distinguish these data points; therefore, their paths are shorter in the trees. The nodes colored green indicate that these are normal data points requiring more splits to separate and thus a lengthier path length. The red-colored nodes indicate the outliers-isolated early in the tree-this is indicative of a shorter path.

Scores for Outliers: The score is a measure of how easily a point was separated from the others**.** A high value close to 1 (red) means the point is an outlier. Generally lower scores towards 0 are more likely to be part of normal data distribution and depicted in green. In the

structure, outliers with smaller path lengths as shown for the red nodes are indicated for additional scrutiny or removal, whereas points which have larger paths, like the green nodes, are part of the normal dataset. Outlier Detection Process In the proposed framework, this can be achieved by utilizing the Isolation Forest method that aids in isolation of outliers in well log data by producing several such trees.

The average path length is obtained by passing each of the data points through several trees, and this is used to calculate the anomaly score. Data points are flagged as possible outliers if they were consistently isolated early in trees or, in other words, with short path lengths. In the image, red nodes are these outliers that have high anomaly scores in other words they vary a lot from most of the data. In short, the image indicates how Isolation Forest picks out outliers by building a set of trees and measuring how much path is needed to separate each data point. The smaller the distance measured, the more probable the point is that it is an outlier. This aspect is one of the most important in the proposed framework as it efficiently picks out or separates the outliers in big well log datasets that are high dimensional, thereby improving the quality and accuracy of the machine learning models.

3.4.4 COMPARISON OF OUTLIER DETECTION METHODS FOR WELL LOG DATA

Comparison of Three Representative Methods Most Adopted: One-class SVM, Isolation Forest, and LOF Based on the different concept in underlying approach comparison of three most widely adapted outlier detection methods for well log data is described in Table 9, all based on three algorithms that are supported by a different concept and approach. Such comparisons are made between the key features relevant to detecting outliers in well log data, a complex and high-dimensional domain. The nature of every algorithm can be explained in terms of the types of methods. One-Class SVM is a boundary-based method, while Isolation Forest is based on an ensemble-based method, and LOF works on a density-based approach. In the concept section of the algorithm, understanding how each technique identifies outliers includes. LOF compares the local density of data points to identify outliers. In terms of dimensional handling, One-Class SVM and Isolation Forest are appropriate in high-dimensional environments (Chowdhury, Tanjil, Akter, Amin, & Pal, 2019) (Franco, Corona, AlMahroos, Kestring, & Oliveira, Practical guide to successful wireline formation pressure testing across highly deviated and depleted sections, 2022); LOF is exceptional where the data environment has varying densities. The computation efficiency of Isolation Forest is at its best because it is of tree-based structure, and the LOF and One-Class SVM will have a more middle

range computational need in the light of factors like selection of kernel and neighbourhood search. Key strengths of each method are also underlined here: One-Class SVM can describe nonlinear relationships, Isolation Forest is scalable and computationally light, and LOF is excellent at local anomalies in datasets with variable density regions, like geophysical well log data. In the column describing the approach to outlier detection, each approach describes how it locates anomalies: one by a definition of boundaries, the other by isolating where the records end up in random trees, and the last one by comparing local density differences. It performs well in the detection of subtle, complex anomalies suitable for well log data; Isolation Forest is the right choice for efficient handling large well log data sets; and it is very effective for finding local outliers within geologically heterogeneous regions. Finally, weaknesses are discussed: for example, One-Class SVM requires parameter tuning, Isolation Forest may miss context-specific anomalies by inherently being random, and LOF is inefficient for global outliers. Anomaly score range is the added detail on how each method outputs its anomaly scores—either in form of a binary distribution or a continuous value or by distributing the density around the data point. In general, this comparison is crucial for proper selection of the outlier detection technique based on the inherent characteristics of the well log data; it is a compromise in between the efficiency of computations, dimensionality, and nature of anomalies to be detected.

Table 9: Comparative Analysis of One-Class SVM, Isolation Forest, and Local Outlier Factor (LOF) for Outlier Detection in Well Log Data

Aspect	One-Class SVM	Isolation Forest	Local Outlier Factor (LOF)
Method Type	Unsupervised, Boundary-based	Unsupervised, Ensemble-based	Unsupervised, Density-based
Algorithm Concept	Constructs a hyperplane to separate normal data from outliers	Isolates outliers based on random feature selection and partitioning	Detects outliers by comparing local density with neighbors
Dimensional Handling	Suitable for high-dimensional data	Efficient for large, high-dimensional datasets	Effective in datasets with varying densities
Computation Efficiency	Moderate (depends on kernel complexity)	High (due to random tree-based structure)	Moderate to High (due to Neighbor search and density computation)

Key Strength	Captures non-linear relationships via kernel functions	Scalable, computationally light, and suitable for large datasets	Effective in identifying local density anomalies, adaptable to variable density regions
Outlier Detection Approach	Identifies data points outside the learned boundary	Isolates points requiring fewer splits in decision trees	Identifies points with lower density compared to neighbors
Suitability for Well Log Data	Excellent for detecting subtle anomalies in complex, high-dimensional data	Well-suited for large-scale well log datasets due to scalability	Effective for detecting local anomalies in regions with heterogeneous geological conditions
Weaknesses	May require parameter tuning (kernel selection, ν parameter)	Relies on randomness; may miss some context-specific outliers	Not suitable for detecting global outliers; relies on local density
Anomaly Score Range	Binary decision boundary (inliers or outliers)	Continuous score (0 to 1, higher indicating more likely an outlier)	Continuous score based on relative density (higher values indicate outliers)
Application Context	High-dimensional, complex datasets where capturing non-linear relations is critical	Large-scale datasets requiring fast computation and scalability	Datasets with varying densities or local structures, such as geophysical well logs

3.5 FEATURE SELECTION USING VARIOUS METHODS

Feature selection is the most critical phase of the data pre-processing procedure in any machine learning model, especially when handling high-dimensional datasets, such as in borehole logging data, which is often composed of several measurements and variables. Its aim is to retain only the most significant informative features for the ML model, simplifying the complexity and raising predictability accuracy. A large number of measurements are taken across several parameters when using the well-logging data but each will be contributing differently to the predictability of reservoir properties. Not all of them carry equal importance, though- removal of irrelevant or redundant features may make the model more efficient,

prevent overfitting, and make results more interpretable. In such cases, the model learns to see the patterns in the training data that may not fit well to the unseen data. This essentially leads to poor performance. Thus, the role of feature selection becomes significant in the design of good predictive models. In this system, the three statistical techniques most widely used are:

- **Correlation Coefficient**
- **Principal Component Analysis (PCA)**
- **Variance Threshold Plot**

All these techniques are being used for feature selection. These techniques will help us figure out the most important features, given that they significantly contribute to predicting the target variable, filtering out the rest because they either just add noise or redundancy. The correlation coefficient estimates the linear relationship between the features. This way, I can remove one of two features, which are significantly correlated with each other. Principal Component Analysis (PCA) simply reduces a high-dimensional dataset by choosing a smaller set of uncorrelated principal components that account for most of the variance in the data. Finally, the Variance Threshold Plot eliminates features with very little variance, since those features would not contribute much toward the prediction task. Together, these feature selection techniques reduce the dataset which makes the ML model more precise, simpler, and more interpretable.

3.5.1 CORRELATION COEFFICIENT

In logging boreholes, such a coefficient is critical in portraying how different well log measurements are associated with one another. The ability to determine which features to retain for use in machine learning models built to predict reservoir properties depends upon understanding these relationships. By finding the correlation coefficient between some variables, I can measure if some attributes are redundant or complementary, which will guide the selection process in the best way to maximize model performance. The most commonly used metric is the Pearson correlation coefficient. The Pearson correlation coefficient is the most commonly used metric for this purpose. The formula for calculating r_{xy} is given in the eqn (10):

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}} \quad (10)$$

In this equation:

r_{xy} represents the correlation coefficient.

$x_i$and y_i are individual sample points for x and y.
$\bar{x}$ and $\bar{y}$ are the means of x and y, respectively.
n is the number of observations (samples).

The above formula results in a number between -1 and +1. The closer that a correlation coefficient is to +1, the stronger the linear relationship tends to be, meaning that increases tend to go with increases, and decreases with decreases. A correlation coefficient close to -1 means very strong negative linear relationship in which increase of one variable is accompanied by a decrease in another. The coefficient close to 0 signifies that one can say there are no significant linear relationships between two variables. In this model, the correlation analysis suggested that Sonic log (AC) was highly positively correlated with the other feature of particular interest: Density (DEN), Neutron Porosity (NEU), and the Photoelectric Factor (PEF). The correlations are meaningful because these features are physically interrelated with each other; Sonic log measures the speed of sound through rock-a quantity inherently influenced by porosity and density. Therefore, retaining these highly correlated features will be important toward the proper prediction of reservoir properties as they each convey extra information to the model. On the other hand, features like RMED (Medium Resistivity), RDEP (Deep Resistivity), and Gamma Ray (GR) were proven to have low correlations with AC, DEN, and NEU. These weaker correlations suggest that these features are not at all correlated with the target variables and may be noise in the model, which degrades its predictive performance. As such, the procedure removed systematically features with weaker correlations from the datasets. This selective procedure streamlined the modeling workflow such that one could maximize the most relevant and information-rich features in preference to less significant ones, thus enhancing model accuracy and efficiency. Consequently, the modeling process becomes more focused by removing the less correlated features, hence reducing complexity with a superior quality of prediction. This procedure ensures that a relevant subset of the features selected contributes significantly to the model without incurring redundancy-an essential requirement for both optimizing model performance and interpretability.

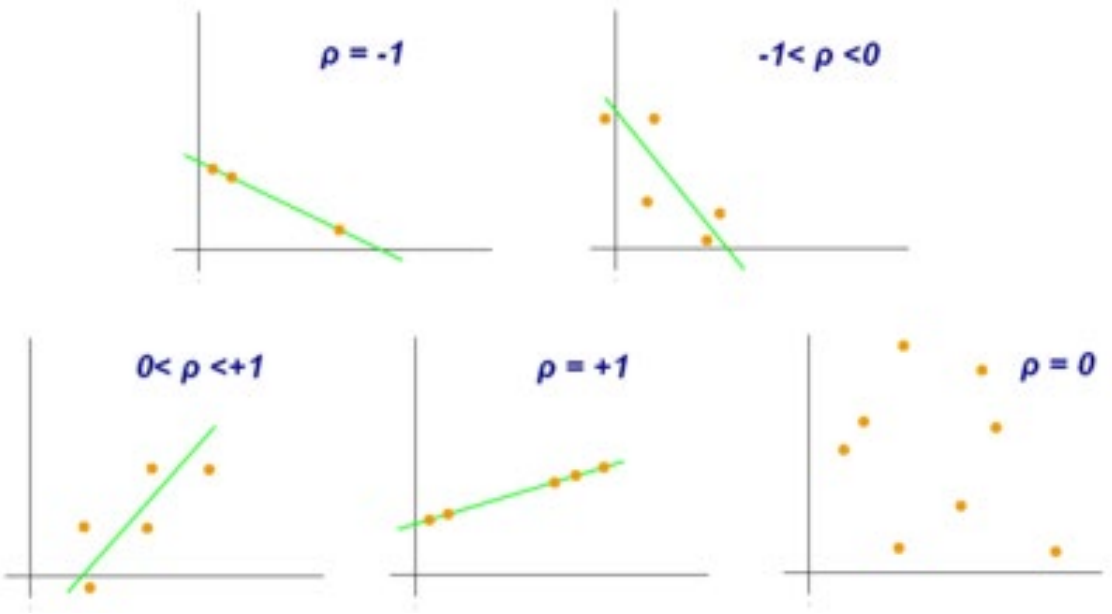


Figure 25: Feature selection using Correlation Coefficients

Figure 25 describes several scenarios by the values of the correlation coefficients, which are very essential to interpret the relationship that can be involved between two variables in a dataset.

- **ρ = -1 (Perfect Negative Correlation)**

In this case, the two variables have a perfect negative linear correlation. When the level of one variable goes up, the other will go down in an absolutely predictable manner. All the points on the graph lie exactly on a line sloping downward such that every increase in one variable results in a fixed, predetermined decrease in the other.

- **-1 < ρ < 0 Negative Correlation**

This scatterplot displays a negative correlation where the changes in one variable often lead to reductions in the other variable. Yet, the pattern does not describe a relationship that is linearly consistent. The dots do slope down but are farther apart compared to the previous illustration. Therefore, though there is a negative relationship present, it is not closely predictable.

- **0 < ρ < +1 (Positive Correlation)**

In this case, the correlation is positive, so if one of the variables increases, the other variable also tends to increase. Again, the points are not on a line, so the relationship seems to vary some. The trend line is sloping upwards, which indicates the positive correlation.

- **ρ = +1 (Positive Correlation or Perfect Positive Correlation)**

In this case, both the variables are perfectly positive correlated. The one variable increases, the other also increases in a perfectly linear relationship. The points lie perfectly upon an upward sloping line, which shows a one-to-one correspondence between the variables.

- **ρ = 0 (No Correlation)**

In this case, the two variables are not in any linear relation with each other. The points are scattered irregularly across the graph without any trend or pattern being recognized. Hence, an increase in one variable does not predict the same change in another variable. Therefore, the two must be independent of each other.

Over the general consideration, the coefficient of correlation is one of the highly important feature selection metrics for science applications that include machine learning and data analysis. Knowing how the different features are correlated helps practitioners to identify the redundant and complementary variables leading to effective modeling. Features highly correlated with the target variable are usually retained; where weak or no correlation exists, they may sometimes be excluded towards simplification of the model and, thereby, achieving better predictive performance.

Feature Selection Process:

Step 1: Calculate the correlation matrix of the features present in the dataset.

Step 2: Assume the threshold value as 0.9. Remove one of the two features if the absolute correlation value surpasses the threshold value.

Step 3: The remaining low inter-correlation feature contains the lowest redundancy.

The correlation-based feature selection process follows a step-wise approach to construct models with less redundancy among the features to increase the performance of a model. It computes a correlation matrix for all the features in the dataset. This computation quantifies the pair-wise correlations between each feature. Then, it sets a threshold value, usually at around 0.9 and checks for high correlation between features. If any two features show a correlation above this threshold value, one of those features is ruled out from being considered. Step 3 will produce low dimensional features and also has low inter-correlations-so as to reduce redundancy while making sure that this model is retaining only the most informative variables. This technique de-clutters the dataset and reduces overfitting chances and also enhances the interpretability of predictions produced by this model. The focus of this model is on getting

those features that impart unique information while improving the accuracy and efficiency in prediction.

3.5.2 PRINCIPAL COMPONENT ANALYSIS

Principal Component Analysis or PCA is one of the key techniques for reducing the dimensionality of the data and is widely used in statistics and machine learning. In nontechnical terms, PCA transforms a bunch of correlated variables into a few linearly uncorrelated variables, known as principal components, retaining as much of the original variance. This transformation is crucial when handling large sets of data, such as those found in borehole logging, since numerous features often add up to overlapping information regarding reservoir properties. The mathematical definition of this transformation is provided by the eqn (11):

$$PC_k = \sum_{i=1}^{n} w_{ki} x_i \tag{11}$$

Where:

PC_k denotes the Kth principal component.

w_{ki} is the weight for feature i in component k.

x_i represents the original features.

In the context of the presented predictive framework of reservoir properties, PCA plays a critical role in the identification and preservation of variables responsible for most of the variance with the given data set. It has been determined that Sonic (AC), Density (DEN), and Neutron Porosity (NEU) explain a large portion of the dataset's variance and are considered prime candidates for further modeling efforts. Through focusing on influential variables, PCA increases the performance of any model while simultaneously decreasing the complexity of the analysis. It enables researchers and data scientists to interpret the results more effectively by narrowing the focus toward a smaller subset of features that holds most information about the underlying geological formations.

The PCA process incorporates some critical steps: first, data standardization; a dataset is centered on the mean and scaled so that each feature contributes equally to the analysis. Second, after standardizing the features, the covariance matrix of features is calculated to evaluate all possible inter-correlation features. Then the covariance matrix is used to calculate eigenvectors and eigenvalues, where the eigenvectors represent directions with maximum variation and the

eigenvalues indicate the magnitude of this variation. The transformation can be summarized in the following formula:

$$Z = XW \tag{12}$$

Z is the data matrix of principal components. (new features)

X is the original data matrix

W is the matrix of eigenvectors. (also known as principal components).

The principal components retaining the highest variance are those eigenvectors of the largest eigenvalues. The following illustrates the variance of each principal component by its corresponding eigenvalue:

Variance:

$$Variance(PC_i) = \lambda_i \tag{13}$$

Where

λ_i is the eigenvalue corresponding to the $i - th$ principal component.

The principal components that possess the largest eigenvalues are retained. This is because they add up to sum the most variance in the data. The variance of each principal component is measured by the eigenvalue λ_i, which gives a measure of how much that component may be used to explain the variability in the dataset. These strict selection criteria will remove features with minimal contributions to variance—such as RMED, RDEP, and GR—which will thus streamline the dataset and prevent noise from potentially degrading the model accuracy.

Applicaton of PCA makes the modeling procedure much easier by reducing the risk of overfitting since the models can be found usually overcomplex in case irrelevant features are considered. Attention to principal components will increase the interpretability of the results, which will be clearer when expressed in terms of the real geological properties of the reservoir. Finally, PCA provides for better predictions and reliable accuracy in reservoir characterization and presents an avenue for better decision making in the adopted hydrocarbon recovery strategies, and it ensures that the most relevant variables informing the subsequent analyses are used effectively.

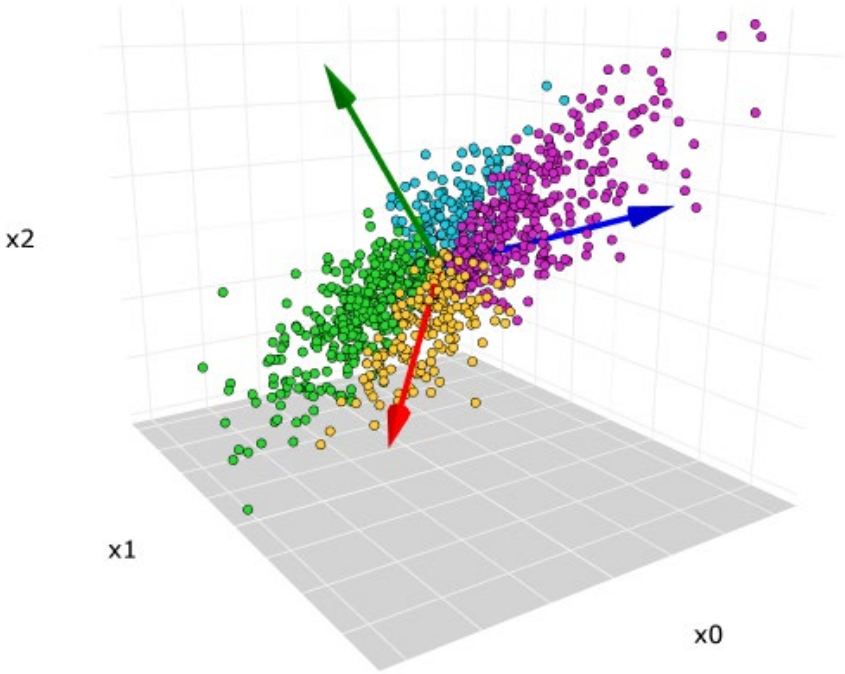


Figure 26: Feature selection using PCA

The concept of PCA in three dimensions is shown in the following Fig. 26 where the points' distribution as well as how they relate with the principal components have to be visualized. The original features or dimensions of the dataset are axes x_0, x_1, and $x_2$with colored points in the space describing individual observations or samples on how they relate with each other multidimensionally. The arrows in the figure represent the principal components created from PCA. Every arrow point in the direction of maximal variance within the data set, meaning the directions along which the data is maximally spread out. This plot emphasizes that PCA really does tend to come up with a good simplification of the data set, which should be the aim when reconstructing a log. This framework proposed the enhancement of the power of machine learning algorithms for the estimation of missing or non-recorded logs by identifying and utilizing the most relevant features explaining a significant proportion of variance in well log data, which will enhance the overall performance and interpretability of the model. In this aspect, PCA is one of the critical steps in an integrated approach toward AI/ML-assisted log reconstruction, as it eventually ensures the most important and informative features are used during prediction modeling.

PCA Feature Selection Process

Step 1: Data standardization is done or mean-centred and scaled.

Step 2: Calculation of the covariance matrix.

Step 3: Find the eigenvectors (principal components) and eigenvalues.

Step 4: Selection of the principal components that capture a specified amount of variance.

Step 5: Plot the data onto the selected principal components.

The feature selection of PCA begins with Step 1 by standardizing the data, that is, centering it around the mean and scaling it so that each feature will contribute equally to the analysis. In Step 2, it computes the covariance matrix of the standardised data in order to assess how the variables co-vary with each other. In Step 3, it will calculate the eigenvectors and eigenvalues of this covariance matrix; these eigenvectors are direction vectors in which the maximum variance lies, and the eigenvalues describe how much that variance is. To do this, in Step 4, we retain principal components that correspond to the largest of these eigenvalues, meaning those that explain most of the variance in the data set. Finally, Step 5 is to plot the actual data on the selected principal components. Thus, one gets new features that contain most of the information of the original data but the dimensionality is reduced. This helps in enhancing model efficiency and interpretability as well as minimizing noise and redundancy in the data set.

3.5.3 VARIANCE THRESHOLD PLOT

The threshold variance drops all of those features where their variances are less than a defined value, meaning there is not enough useful information for the prediction. So, I shall show the mathematical formulation. The Variance Threshold Plot is important in feature selection processes especially in contexts involving high-dimensional data like that which is often found in the well logging analysis. This method is used to enhance the quality of the dataset by removing all those features that do not add meaning to the predictive modeling process. The underlying mathematical formulation for calculating the variance of a feature X_j is expressed as in the eqn (14):

$$Var(X_j) = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x_j})^2 \tag{14}$$

Where:

$X_j$is the $j-th$ feature.

x_{ij} is the $i-th$ observation of feature X_j.

$\bar{x_j}$ is the mean of feature X_j.

n is the number of samples.

From the variance computed for each feature have; we can then understand the spread of the data. Low variance means that the values of a feature are only lumped around its mean; thereby the feature is of little use for prediction purposes. Feature selection begins with the computation of the variance for each feature in the dataset. Next, a threshold for variance is set, say 0.1, and this now becomes the filtering criterion in which those features that do not have much variance, or whose variance is below this threshold, are removed from the data set, which is least likely to provide much information about the underlying patterns. This step is necessary to remove noise in the data set and reduce the chance of overfitting, where a model can "catch too much detail" and not the actual patterns in the data

The leftover set of features from applying the variance threshold shall have variance greater than that set in the limit. Such features are more likely to be informative and meaningfully contribute to the efforts of predictive modeling. It helps in this regard by filtering out the dataset in such a way that Variance Threshold Plot enhances the overall performance and interpretability of the machine learning models as an integral part of AI/ML-assisted log reconstruction to estimate missing or non-recorded logs. This approach provides a reduction of complexity in the model formulation and enhances prediction robustness based on emphasis on most relevant and informative features.

Feature Selection Using Threshold Plot Process

Step 1. Calculate variance for every feature.

Step 2. Set a threshold variance, say, 0.1. Features that have a variance lower than that will be eliminated.

The remaining features will have a variance greater than the threshold determined in this step, and therefore it follows that they are worth considerable spread and a good probability of being far more informative to the model.

Feature selection using variance threshold has the following three steps devised in an attempt to improve the quality of the data to use in predictive modeling. The first step is to calculate the variance for each feature within the dataset, which computes how much values of that feature diverge from the mean. Then, based on variance, a threshold can be selected, such as 0.1; the point of this threshold is to filter features. Those features whose variances turn out to be less than the threshold is less informative and are discarded from the data set. Remaining features that have a variance greater than threshold imply significant spread in the value, and

therefore can give useful information for training models. In this manner, it filters noise and redundancy, and thus makes models better and more interpretable. Finally, it helps refine the feature set to be more relevant to the attributes for application use in outcome prediction such as well logging analysis.

3.5.4 FEATURE SELECTION USING CORRELATION COEFFICIENT, PCA, AND VARIANCE THRESHOLD PLOT

feature selection procedure with three important methodologies focused, namely Correlation Coefficient, Principal Component Analysis (PCA), and Variance Threshold Plot. The first step of the diagram starts at Raw Data that is a set of data gathered from borehole logging, containing several measurements taken in these points. All these measurements have various magnitudes of relevance to the target property, such as in the characterization of the reservoirs. The first and foremost purpose of this diagram is to represent how these three techniques work to choose the most informative features and discard the redundant or irrelevant ones. The correlation coefficient approach computes a correlation matrix for all of the features in the dataset. The correlation coefficient is the measure that determines both the strength and direction of the linear relationship between a pair of variables. This correlation matrix is actually an analysis tool that describes the interactions in the various features. A threshold is then put on the correlation coefficients, and if the correlation of any feature pairs exceeds this threshold, then one of the correlated features will be removed. This step now takes away the redundancy that would make the selected features duplicate towards building the predictive model.

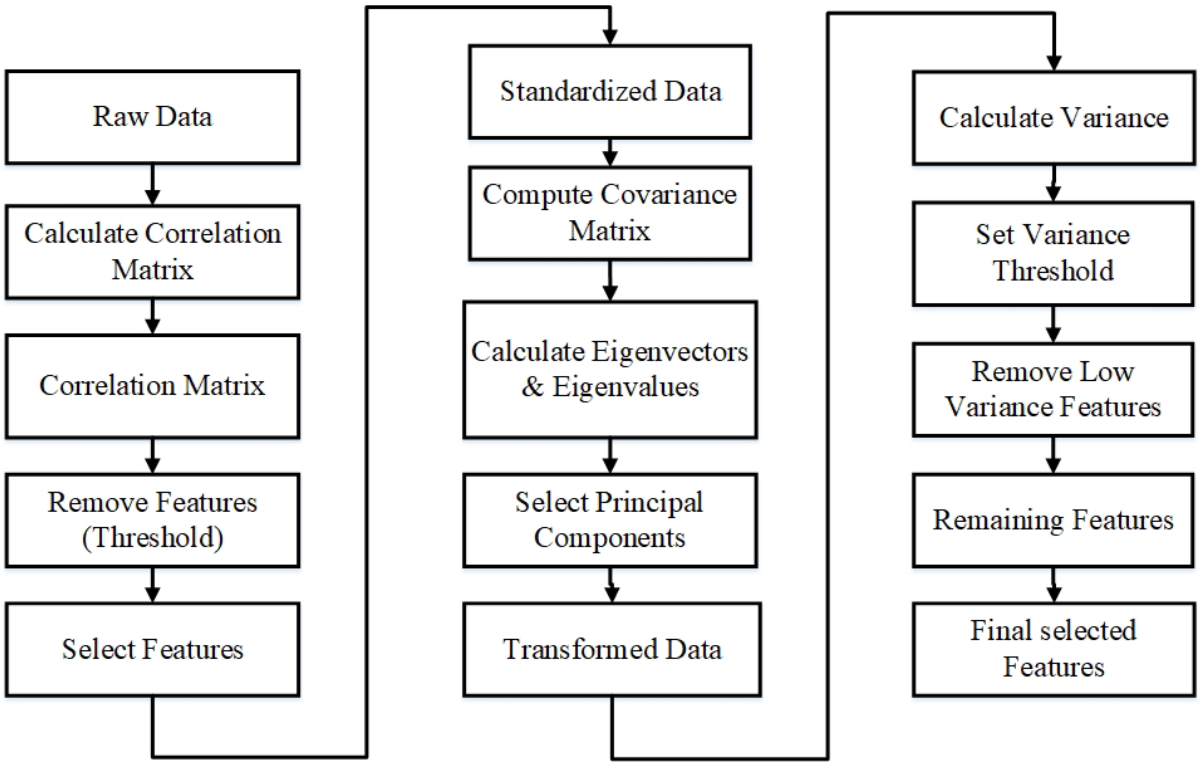


Figure 27: Feature Selection using Proposed Methodology

The diagram then moves on to the trajectory of Principal Component Analysis or PCA. PCA first standardizes the data so that all features are on comparable scales. Covariance calculations are then done on standardized data so that how much the features have varied from the mean with regard to each other is captured. This covariance matrix is used to compute eigenvectors and eigenvalues. These represent the directions and magnitudes of the variance in data, respectively. The newly derived uncorrelated features - called principal components - are chosen based on their eigenvalue, so that those components explain the most variance. This dimensionality reduction simplifies the dataset, keeping only the variables that most impact the target property while discarding those with little contribution to the model's predictive power. The third method applied in the figure is the Variance Threshold Plot. In this method, variance for each feature is calculated, and a predefined threshold, such as 0.1, is set. Features below the threshold value of variance are uninformative and thus removed from further analysis. It is essential because features with low variance are not strong enough to make a better prediction. So, only such features with maximum variance should be retained.

At last, the diagram reaches the Final Selected Features stage where the features selected using each of the three methodologies come together. This integrated approach, for instance, uses the

dataset to streamline only the features which are most relevant; hence further modeling efforts shall only concentrate on such features. The feature selection process can in turn dramatically enhance the predictive performance of the machine learning model with reduced complexities, while minimizing the risks of overfitting and facilitating interpretable results as it combines the three techniques. This systematic selection of features is significantly helpful for high-dimensional datasets such as borehole logging data, where many features may contain redundant or irrelevant information. The resulting set of selected features forms a strong basis for accurately estimating missing or non-recorded logs in the context of the proposed framework, integrated approach on ai/ml assisted log reconstruction methodology for estimation of the missing/non-recorded logs.

3.6 CLUSTERING PROCESS POST-FEATURE SELECTION

As I have selected the features, the next and very important step in our integrated approach to AI/ML-assisted log reconstruction is clustering. In this phase, the data is organized into clusters based on the responses of selected logging measurements from other wells. Clustering will be an important tool in analytics to look for the relations among the different wells for understanding similarities and patterns in general geological grounds. Grouping similar wells would be an effective way of gaining greater insight into characteristics of reservoirs involved in resource management and their exploration. In this direction, towards the execution, we have used Gaussian Mixture Models, or GMM (Yang, et al., 2016) (Wu, Jain, Kulkarni, & Abubakar, 2018) (Tavakkoli & Amini, 2006). The probabilistic model seems particularly appropriate to identify complex, oblongly shaped clusters in data. Unlike other traditional clustering techniques, such as k-means, that assign hard labels to every point, the GMM framework gauges the probabilistic relationship between data points by assigning every sample a probability score of the probability of that particular sample belonging to a certain cluster. This probabilistic framework is particularly valuable in dealing with geological data that hosts intricate underlying structures and which cannot be easily pigeonholed into well-defined categories (Song, 2023) (Rebelle & Lalanne, 2014).

3.6.1 CLUSTERING USING GAUSSIAN MIXTURE MODELS

Within the developed framework titled "Integrated Approach on AI/ML Assisted Log Reconstruction Methodology for Estimation of the Missing/Non-Recorded Logs," the clustering process is more critical in order to carry out the analysis related to the geological characteristics of a reservoir by using borehole log data. The use of GMM is very helpful in getting meaningful clusters from selected responses of the logs (Luo, Sun, Zhang, & Liu, 2024)

(Lee, Byun, Kim, & Yoo, 2017) The next subsection discusses how GMM operates in this context. In the context of this framework, we start by preprocessing before actually clustering using techniques such as correlation coefficients, PCA, and Variance Threshold for selecting the most relevant features that contribute towards borehole log data. These steps thereby preserve only those features that considerably contribute to reservoir attributes predictions and it is along these selected features that the entire process of clustering stands (Mahmoud, Elkatatny, Ramadan, & Abdulraheem, 2016) (Gill, Shomrony, & Fligelman, 1993) (Euzen & Power, 2014).

i. Understanding Gaussian Mixture Model

GMM is based on the presumption that the data is generated by a mixture of several Gaussian distributions. That is to say, every Gaussian distribution can be taken to represent a cluster in the dataset (Emelyanova, Peyaud, Dance, & Pervukhina, 2020) (Silveira, Figueiredo, Lima, Gonçalvez, & Amaral, 2024) (Cherana, Aliouane, Doghmane, Ouadfeul, & Nabawy, 2022). GMM allows clusters of different shapes and sizes, which makes it particularly suitable for complex geological datasets in which the relationships between the various log measurements may not be spherical. The mathematical basis of GMM is supposed to be as follows: data points are distributed in a mixture of Gaussian distributions. It can be represented by the following probability density function for the mixture model as it was given in the eqn (15):

$$P(x) = \sum_{k=1}^{k} \pi_k \aleph(x|\mu_k, \sum k) \tag{15}$$

Where:

$P(x)$ is the probability density function for the data point x.

K represents the total number of clusters.

π_k is the mixing coefficient for cluster k

$\aleph(x|\mu_k, \sum k$ denotes the Gaussian distribution with mean μ_k and covariance $\backslash\sum k$ for cluster k.

The Gaussian distribution itself can be defined by the eqn (16):

$$\aleph(x|\mu_k, \sum k) = \frac{1}{(2\pi)^{d/2}|\sum k|^{1/2}} \exp\left(\frac{-1}{2}(x-\mu_k)^T \sum_k^{-1}(x-\mu_k)\right) \tag{16}$$

Where:

d is the number of dimensions (i.e., features)

$\sum k$ is the determinant of the covariance matrix.

The GMM algorithm essentially works by performing an iterative process of refining the parameters k in order to maximally describe the observed data under the model. In general, that is achieved by applying the Expectation-Maximization (EM) algorithm. It comprises two major steps: The Expectation Step (E-Step) and the Maximization Step (M-Step) (Bagheripour & Asoodeh, 2013) (Alizadeh, Najjari, & Kadkhodaie-Ilkhchi, 2012) (Ali & Zhu, 2023).

1. E-Step: It computes the responsibility that every data point x_i has in belonging to the cluster k as followed by the eqn (17):

$$\gamma_{ik} = \frac{\pi_k \aleph(x_i|\mu_k,\sum_k)}{\sum_{j=1}^{K} \pi_j \aleph(x_i|\mu_j,\sum_j)} \tag{17}$$

Where γ_{ik} represents the probability of data point x_i belonging to cluster k. This assignment helps in understanding how likely each sample is to be a part of the identified clusters (Ali & Sheng-Chang, 2020) (Albuslimi, Alkalby, & Al-Taweel, 2021) (Al-Mudhafar, Lawe, & Noshi, Clustering analysis for improved characterization of carbonate reservoirs in a southern Iraqi oil field, 2019).

2. Maximization Step (M-Step): After the E-step, the M-step updates the parameters depending on the responsibilities calculated in the above step. The updating formulas for the mixing coefficients, means, and covariances are below in the eqn (18-20):

$$\pi_k = \frac{1}{N} \sum_{i=1}^{N} \gamma_{ik} \tag{18}$$

$$\mu_k = \frac{\sum_{i=1}^{N} \gamma_{ik} x_i}{\sum_{i=1}^{N} \gamma_{ik}} \tag{19}$$

$$\sum_k = \frac{\sum_{i=1}^{N} \gamma_{ik} (x_i - \mu_k)(x_i - \mu_k)^T}{\sum_{i=1}^{N} \gamma_{ik}} \tag{20}$$

Where N is the total number of data points. This process will be iterated till convergence, so that the parameters not change significantly.

ii. Determining the Optimal Number of Clusters

To find the appropriate number of clusters for the GMM, we use statistical selection criteria, namely BIC and AIC. They are statistical models that have emerged in the determination of the appropriate number of clusters that can best elucidate the underlying structure of the data.

AIC is given by the eqn (21):

$$AIC\} = 2k - 2\,ln(L) \quad (21)$$

Where:

k is the number of parameters in the model.

L is the likelihood of the model.

Conversely, the BIC is expressed as in the eqn (22):

$$BIC = ln(N)k - 2\,ln(L) \quad (22)$$

Where N is the count of data points. AIC and BIC estimate model fitting with a penalty for higher complexity. Lower AIC and BIC, the better will be the fit of the model. Hence, we can better decide on an appropriate number of clusters. Using AIC and BIC to decide upon the optimal clustering thereby means an efficiently conducted statistically sound approach that will unearth the inherent patterns within the data. The crucial role of the clustering process with GMM is reflected in making well-specific clusters using selected logging measurements. More importantly, this improved approach towards clustering very much enhances our capability to analyze reservoir characteristics (Akinyokun, Enikanselu, Adeyemo, & Adesida, 2009) (Puskarczyk, 2020) (Abbas, Al-Mudhafar, Anees, & Wood, 2024). Advanced statistical criteria like BIC and AIC are now being utilized for facilitating the determination of the optimal number of clusters for achieving increased reliability and interpretability of the outcomes from our clustering. Ultimately, this robust method substantially contributes to the overall framework for estimating missing or non-recorded logs, enhancing the precision of predictions developed within the integrated approach of AI/ML-assisted log reconstruction (Torghabeh, et al., 2014) (Abed & Hamd-Allah, 2019) (Riadi, Istiyanto, Ashari, & Subanar, 2013).

3.7 ML PREDICTION USING RANDOM FOREST ALGORITHM

In the proposed framework, the Random Forest algorithm has been used as one of the highly promising predictive techniques for the best estimation of borehole log data. The methodology works well with the inherent complexity and high dimensionality of well log data. In these cases, traditional modeling approaches may have difficulty in capturing detailed relationships between various log responses and reservoir properties. Random Forest is a type of ensemble learning method used primarily for classification and regression tasks in machine learning. It trains many decision trees and then combines the outputs to get the final better prediction. This approach will play to the strengths of decision trees while reversing some of the weaknesses at the same time, such as noise sensitivity and overfitting.

3.7.1 RANDOM FOREST STRUCTURE

The algorithm is based on a decision tree. In the case of the Random Forest, each of the individual trees is in reality an independent learner within the ensemble that contributes to the ultimate output. Some of the key elements and procedures for creating a Random Forest are: Bootstrap Aggregating, or Bagging: An ensemble generated by random forest employs bagging to build an ensemble over different subsets of the training dataset. In such an approach, each tree is trained on a randomly selected subset of the data, drawn with replacement. That means some samples are repeated and others may not be included at all. Introducing variability into the training data helps decrease overfitting, one of the common problems that may be encountered when learning particular patterns from noise inside the training set.

An additional way in which Random Forest operates different subsets of the data is by using randomness to select a subset of features for each decision tree. This process is known as random feature selection and merely provides further guarantees of diversity because each tree is built from seeing the data from a different angle. For instance, if there are p features in the dataset, only a subset of mmm features is used at any node of the tree in consideration, where m<pm < pm<p. It is this randomness that averts the trees from becoming too similar to each other and enables better generalization across the ensemble.

3.7.2 BUILDING DECISION TREE

Each tree in the Random Forest is built recursively as follows:

- **Node Splitting**: At each node, it is examining all possible splits according to selected features; thus, it will go with the one that separates the data the best according to a particular criterion or choice: for example, Gini impurity in classification, and Mean Squared Error in regression.
- **Leaf Nodes**: Once a node reached some criterion, for example, maximum depth, minimum samples per leaf, or a purity threshold, then it would become a leaf node that represents a final output value in case of a regression problem and class label in case of a classification one.
- **Ensemble Output**: The final outputs from all the trees are combined for giving out an output. For regression tasks, the average of the prediction of each tree is obtained. In the case of classification tasks, a majority vote is taken.

The predictive model of a Random Forest can be mathematically represented by the eqn (23):

$$\hat{y} = \frac{1}{N}\sum_{n=1}^{N} f_n(x) \quad (23)$$

Where:

$\hat{y}$ is the final predicted output for the given input feature vector x.

N is the total number of trees in the forest.

$f_n(x)$ is the output of the $\boldsymbol{n} - th$ decision tree for the input x.

This expression indicates how the ensemble model averages the output of every individual tree to make a final prediction. The averaging of many predictions works quite well in random forests to reduce individual errors and variance, thereby increasing the precision and robustness of predictions.

- **Reduced Overfitting**: Reduction overfitting often occurs due to the Random Forest model; it averages the outputs of multiple trees trained on different subsets of data and features. Individual trees, like many other models for predictive modeling, are often prone to overfitting and thus tend to perform worse when tested on unseen data.
- **Feature importance**. Along with the prediction accuracy, the Random Forest also provides information on feature importance. This could be very useful, for instance, in geology and some log measurements that may influence characterization of the reservoir to some extent with a different degree of significance.
- **Robustness to Noise**: The ensemble approach improved the robustness of the model to noise and outliers in the data set, improving the accuracy of predictions made by the model. Different trees might capture the different aspects of the underlying patterns in the data, thus giving a more holistic view of relationships among variables.
- **Scalability**: Random Forest can accommodate large datasets with many features without much loss in performance, thus suitable for more complex datasets characteristic in borehole logging and geological analysis.

Random Forest algorithm, in summary, is a robust and versatile method when tackling prediction tasks in machine learning. This framework can attain high accuracy with generalization by reducing the risk of overfitting through the construction of multiple decision trees using bootstrap aggregating and random feature selection. Also, being a framework averaging the output of numerous trees, the mathematical framework hardens its efficacy, making the above approach very popular in lots of applications, such as just estimating missing

or nonrecorded logs in geological studies (Gupta & Shah, 2022). Hence, it forms an integral part of the combined strategy of AI/ML-based log reconstruction techniques and hence improves the robustness as well as interpretability of models used in making predictions in reservoir characterization

3.7.3 HYPERPARAMETER TUNING AND CROSS-VALIDATION IN RANDOM FOREST

Hyperparameter tuning is an important process in that it directly impacts the performance and ability of generalization of the Random Forest model. Proper tuning of such hyperparameters may make a huge difference in model improvement and can avoid problems such as overfitting or underfitting.

Some of the most important ones include

This is the number of decision trees constructed in a forest. Generally, the more trees one uses, the greater the stability and accuracy of predictions because the variance in the model has been lowered. On the other hand, the more trees mean longer computation time. In an ideal scenario, such balance should be struck to find the right combination between performance and efficiency.

- **Maximum Depth**: The number of splits the tree is allowed to do before it thinks it's complete. The more deeply a tree is allowed to grow, the more complex relationships in the data it can capture, but overfit- to detail in the noise of training data. In practice, the depth of trees may be limited (early stopping criteria). This actually simplifies the structure of the model and enables its generalization.
- **Minimum Samples per Leaf**: This hyperparameter is the minimum number of samples that have to appear in a leaf node, which is the final decision point of the tree. A higher value ensures that each node represents a good proportion of data points so that overfitting doesn't occur. This will make the model focus more on the robust patterns in data, which boosts its ability to generalize new data points.

3.7.3.1 Cross Validation Procedure

Cross-validation is a good robust technique that measures the ability of the model to generalize for data which have not been seen and to prevent overfitting.

Steps of Cross-validation:

- **Dataset Partitioning**: In k-fold cross-validation the dataset is partitioned into k equal size subsets (for instance, k = 5). Each fold is used as validation set once and the remaining k-1 folds are used to train the model. In this proposed framework, fivefold cross-validation has been used which means one subset is used as validation and the remaining four as training.
- **Training and validation Cycles**: The model is trained on k-1 folds in the training set for each fold; the rest is used for validation. This process is repeated kkk times, and the performance metrics are calculated after each validation set.

The primary feature of a performance metric calculation after each iteration will be the assessment of how well the model has performed on some metrics, such as MSE regression or accuracy in the case of classification, and R-squared in the case of regression. The best from all those iterations is averaged up to give a better estimate as to how well the model generalizes over the whole data set.

The Random Forest model is not overfitting to specific parts of the data and, in fact, working well across different subsets of it by using cross-validation. This prevents us from developing a model that excels at training but is awful at performance on new data - something that is critical when deploying these models into the real world: for instance, predicting missing well logs

3.7.3.2 Training and Data Preparation

Once the step is confirmed, ensuring the robustness of the model by cross-validation, it will be trained using data from the clustering phase. This can be done by fine-tuning the training process so that this information gathered during the clustering process, which groups similar log responses, may now be used (Riadi, Istiyanto, Ashari, & Subanar, 2013).

- **Cluster-Specific Training:** Feature selection and clustering are employed in these steps to preprocess the actual learning data. The data are clustered based on similarities in the log responses, and the resulting clusters of data present with similar geological features. These clusters are then used in training separate models; hereby ensuring the relationships between the features and target variables localize to the specific cluster.
- **Data Preparation**: After the clustering, the data set was divided into clusters. Input features (X, which included the chosen logging measurements) and their respective target values (y, which consisted of the logs to predict) were prepared for each cluster.

This would then enable the model to learn specific patterns and relationships that apply only in that group of wells, hence the forecasts in that separate group are more accurate.

Cluster-Based Training. The data for each cluster is used for training a separate Random Forest model. The strategy can be framed by the eqn (24):

$$Train(X_{cluster}, y_{cluster}) \Rightarrow RandomForestModel \quad (24)$$

$X_{cluster}, y_{cluster}$ represents the features and target for a particular cluster, respectively. Each learned model specializes in the prediction of logs for wells belonging to its own particular cluster and learns relations unique to that group.

- **Offline Prediction**: The trained models will be saved for offline prediction on missing and outlier data in logs. During time of prediction, the top two clusters are computed for every data sample in the prediction dataset. Each of the data samples is processed by the models of the top two clusters and two predictions are generated.
- **Final Prediction Computation**: The final prediction for each sample of data is computed as the weighted average of two cluster-based predictions using normalized probability of the highest two clusters as weights. This also brings in a further level of precision by incorporating the probability of a sample belonging to more than one of the formed clusters: It was given in the eqn (25):

$$\hat{y}_{final} = \omega_1 \widehat{y_1} + \omega_2 \widehat{y_2} \quad (25)$$

Where:

$\widehat{y_1}$ and $\widehat{y_2}$ are the predictions from the top two clusters.

ω_1 and ω_2 are the normalized probabilities of the sample belonging to the two clusters.

3.7.4 UNCERTAINTY ESTIMATION AND VALIDATION

In addition to giving a final estimate, the Random Forest model also computes uncertainty measures at each point of prediction. Given a comparison of the variance of the predictions both across the trees and in different clusters, one can obtain an estimate of confidence at each of these predictions. This type of uncertainty is very useful for proving or validating the model's predictions in actual geological applications where the presence of missing or faulty log data can greatly affect the outcome. This AI/ML-assisted log reconstruction proposal will, therefore, rely heavily on the Random Forest algorithm of the model when predicting the missing data of the log with good accuracy. Further, cross-validation and cluster-specific training with careful

hyperparameter tuning to the specific well clusters are arranged in such a manner that the model uniquely matches the characteristics of the well cluster, ensuring good prediction performance. Further improvement in accuracy is obtained through the method of weighted averaging of multiple clusters, thereby providing robust and reliable predictions. Thus, the randomness of uncertainty measures would be such that the prediction results become accurate, interpretable, and trustworthy for application. In brief, the ML prediction phase involving the Random Forest algorithm in the framework improves the ability for accurate prediction of missing or non-recorded log data. The approach therefore delivers reliable and interpretable results via ensemble learning, hyperparameter optimization, cluster-specific modeling, and uncertainty analysis. This methodology, robustly validated and statistically sound, has given stakeholders the confidence to make informed decisions in reservoir characterization and management. Reconstructing log data to capture subsurface geology complexities would, therefore, be pivotal in advancing log reconstruction and reservoir analysis contribution in the field, buttressing the reliability of AI/ML-assisted methodologies in geological studies. The amount of statistical sophistication and machine learning strength makes this framework the front runner of contemporary geophysical exploration and evaluation approaches.

3.8 INTERACTIVE VISUALIZATION

Lastly, the interactive visualization is through the log reconstruction framework based on AI/ML assistance whereby reconstructed logs and predictions from the machine learning are intuitive and interactive. At this point, communicating the results of the log reconstruction process needs to be conceptualized in a manner that is easy to grasp for geologists, reservoir engineers, and others who are stakeholders. Users can browse the data, validate forecasts, and comprehend relationships among variables using interactive visualization tools. This provides users with the dynamic dashboards to view reconstructed logs alongside the original data, which can show up clearly areas where reconstructed data is missing, noisy, or an outlier. With these tools, users can filter by well, log type, or specific depth intervals to examine the model's performance more closely across regions of interest.

There is an important reason for this step, as one can communicate the result of the log reconstruction process to the domain experts who may make decisions based on the data thus gathered. By visually interpreting reconstructed logs, experts can assess the accuracy of predictions and the reliability of the model for future predictions. Interactive visualization not only engages the users with the results but also motivates the data scientists to work in collaboration with the geologists to make sure that the reconstructed logs are scientifically valid

for further analysis by the geologists. Interactive nature of the tools also allows adjusting and analyzing in real time, in that one can simulate alternative scenarios or look at the uncertainty of the predictions. At the end, such a visualization process enhances confidence in the reconstructed logs and streamlines the decision-making process and facilitates better collaboration among teams.

CHAPTER- 4

RESULT AND DISCUSSION

4.1 BACKGROUND

This AI/ML-assisted log reconstruction framework results section derives to demonstrate the effectiveness and accuracy of the method applied on missing log estimation, outlier detection, feature selection and prediction of geophysical well logs. The overall idea was to improve the quality and completeness of subsurface data through advanced machine learning techniques such as RF and clustering approaches like GMM for robust log prediction and data QC. Log data from wells are one of the most important information pieces regarding characterizations in subsurface geology studies. However, missing values, outliers, and noisy data usually degrade the quality of the analysis in log data. Proper data cleansing, transformation, outlier detection, feature selection, and predictive modeling are thus key to confident outputs. In this study, the combination of different models and algorithms of machine learning across the workflow served to bring good quality and completeness of data.

The mentioned key algorithms for outlier detection were Local Outlier Factor, Isolation Forest, and One-Class Support Vector Machine. It was discovered that LOF best produces geologically representative results on the dataset. Feature selection utilized techniques like Boruta, Correlation Coefficient, Principal Component Analysis, Lasso, and Variance Threshold to minimize redundancy in the data and emphasize the most relevant features. Using GMM-based clustering, wells were grouped based on their log responses, providing better predictions via cluster-specific training for Random Forest models. Based on log reconstruction by the Random Forest, it became possible to illustrate how MLs can robustly predict missing log values based on the information provided. In this context, cross-validation was used in order to evaluate the strength of the model performance while uncertainty was determined using predicted values and comparison to logs originally submitted. The results clearly reflected the success of this integrated framework in achieving high accuracy and efficiency in log reconstruction while minimizing uncertainty.

Results from the AI/ML-assisted log reconstruction framework show significant advances in terms of data pre-processing, outlier detection, feature selection, clustering, and machine learning predictions on geophysical well logs. The framework is capable of surmounting

significant challenges such as missing data, anomalous log values, and noisy signals within the well log data to ensure more reliable and accurate interpretation of subsurface properties.

Key conclusions:

- **Outlier Detection**: Outlier detection algorithms LOF performed the best as an outlier detection method compared to Isolation Forest and One-Class SVM. LOF can easily determine local outliers without any kind of assumption with respect to data distribution-that is, more geologically accurate.
- **Feature Selection**: The methods for feature selection like PCA, Boruta, and Lasso have significantly contributed towards the reduction of complexity in the data by maintaining only the most relevant features that were chosen for clustering as well as prediction purposes. Strong correlation between selected features with the target variables helped increase the accuracy of predictions carried out by the Random Forest model.
- **Clustering**: the use of GMM for clustering provided a probabilistic approach for the log response data aggregation of wells. This clustering step also localized the training that can occur using Random Forest models for better prediction on each cluster.
- **Random Forest Predictions.** I took the Random Forest model after clustering and applied it to predict missing logs. This was done successfully with predictions based on the data that is available. The ensemble learning methodology together with the hyperparameter tuning and cross-validation ensured a robust model that showed greater predictive power.
- **Uncertainty Estimation**: The uncertainty prediction had been almost on par with the uncertainty that is present in the original data acquisition process, which further reflected the robustness of the expected logs.

In a nutshell, integration of machine learning techniques with interactive visualization tools provides a solid foundation for log reconstruction, outlier detection, and data quality control. Thus, results from this research offer a scalable solution that could find applications in several fields of geoscience, enhancing decision-making processes and better interpretation of geological and reservoir studies.

4.2 ANALYSIS OF DATA PRE- PROCESSING

Data pre-processing is a critical step toward ensuring that datasets are given to machine learning and predictive models integrated with integrity and reliability. It encompasses techniques designed to remove errors, inconsistencies, and any irrelevant data points, which would compromise the outcomes of analytic procedures. This chapter outlines some of the most important pre-processing operations-including clipping, detecting anomalies, deletion of constant value columns, and use of logarithmic transformations-which promote higher quality in data. By adopting these methods, the data set becomes more robust and even more suitable for subsequent analysis of data. Hence, prediction performance also improves. Additionally, an AI/ML-based interface allows smoother cycles of quality control, cleaning, and transformation. A new platform is, therefore, facilitated by this new domain that will allow data scientists and geologists to maintain high-quality datasets that will enhance accuracy in geological interpretation and finally in decision-making.

4.2.1 EVALUATION OF CLEANSING AND TRANSFORMATION

The pre-processing stage of the dataset is quite critical for ensuring data integrity and aptness for downstream analysis or modeling workflows. These include the elimination of errors, inconsistencies, and irrelevant data that could compromise the analysis or the results from predictive modelling. Here, a series of significant pre-processing techniques had been applied, including data clipping, removal of anomalous log values, constant continuous value filtering, and logarithmic transformations, all of which added up to enhance the quality of the data.

- **Data Clipping**

Data clipping is a technique to limit the recorded data values within physically meaningful ranges. This is particularly important when sensor or tool malfunctioning causes extreme readings. For instance, in log types like Gamma Ray (GR) or Bulk Density (RHOB), spiking occurs mainly because of errors and never constitutes geologic reality. This clips the values to the boundary so that values fall within an acceptable, realistic range-to say that RHOB values would between 1.8 and 2.9 g/cm³, thus removing outliers from skewing subsequent analysis.

- **Detection and removal of outlier log values**

Outliers are log values far removed from a plausible trend by way of equipment or reading issues. For this purpose, LOF, SVM, and IF algorithms are utilized for detection of these anomalies. These techniques compare each data point with its neighbors to identify the

deviation from the anomaly. For example, spikes or dips in logs such as Resistivity (RT) or Neutron Porosity (NPHI) may become error in data collection rather than geological anomalies. The elimination of these spurious values is important to the maintenance of integrity in the dataset because they may interfere with the development of patterns or trends during training and interpretation using ML.

- **Constant Continuous Values Removal**

There are often long intervals in well log data where specific variables do not change in well logs, especially in homogeneous formations. For example, the Sonic (DT) log or Neutron Porosity (NPHI) is often very stable over one or more long intervals. Such constant values have little to no meaning for modeling purposes and sometimes introduce unnecessary noise, particularly to machine learning schemes that rely on changes from variations to spot trends.

- **Logarithmic Transformation**

Some geologic features, like Permeability (PERM), are typically rather spread across several orders of magnitude, so applying a logarithmic transformation on these data points actually normalized them, thus large values do not dominate the predictive model: the transformation compresses the values into a more manageable range, thus also making more linear relationships among input variables and easier to be interpreted by machine learning algorithms.

The logarithmic transformation helped to improve the data as follows:

Variance Stabilization: The variances stabilized throughout the data set, making it easier for the models to pick through underlying patterns and trends and predict this accurately. Balanced Data: This brings a reduction in the weightage given by extreme values; therefore, an algorithm would be able to realize the geological trends better instead of being skewed by large outliers.

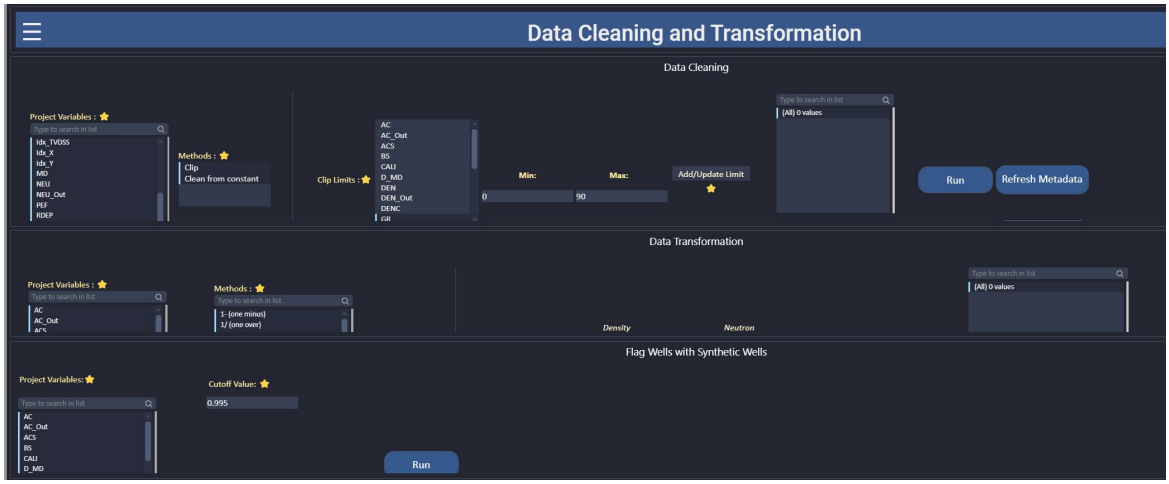


Figure 28: AI/ML Based Data QC, Cleaning and Transformation workflow.

Figure 28 elaborates an interface for performing data quality control (QC) cleaning and transformation workflows using AI/ML techniques. The interface provides the interpreters and data scientists with an interactive user-friendly platform to preprocess well log data before feeding it into predictive models. Below is a detailed explanation of the key components:

- **Project Variables**

The interface brings in all related variables or logs that are used within a project on the left side of the screen. Such logs include Gamma Ray (GR), Bulk Density (RHOB), Resistivity (RT), and Porosity (PHI). Putting all the variables for a project in one view means that interpreters do not need to spend much time figuring out which logs are available for evaluation and which might need cleaning.

- **Methods**

There is the middle part of the interface, where different data cleaning methods are present. These include removing constant continuous values, transformation by log, or removing outliers. Each method is applied as required by the specific needs of preprocessing. For example, users can apply log transformations to values of permeability or choose LOF outlier detection algorithms to remove anomalous points.

- QC Limits

Another important feature of the interface is that one may define QC bounds on individual log variables. One might want to specify what values for each log are acceptable, or one might, for example, want the GR values between 0 and 200 API. If data points fall outside of these bounds,

the interface will automatically flag and correct these anomalies by clipping or flagging them as an error for further scrutiny.

- **Data Cleaning Options**

This section of the user interface allows to launch data cleaning procedures based on the selected methods and QC limits. For instance, a user can select specific limits for Bulk Density (RHOB) and eliminate deterministic values for Sonic Logs (DT). Once all the cleanup methods have been selected, a user can apply them all together to the entire dataset to ensure that all logs are cleaned uniformly.

- **Buttons and Functionality**

The interface provides several action buttons that include adding the QC Limit, which would add custom quality control limits for each variable so that in the handling of data, there could be more flexibility. They also offer the Refresh Metadata, by which users can refresh the metadata in the project and consequently make it relevant with the latest logs along with their properties in the analysis.

- **Run**

This shall execute the entire cleaning and transforming process using any of the selected methods and QC limits. At this stage, the processed dataset is ready to be further analyzed by clustering or machine learning modeling.

Such a result from preprocessing unambiguously shows how much quality has been improved in the dataset.

With application of clipping, anomaly detection, removal of constant values, and logarithmic transformations, the dataset became noise-free, devoid of extreme values, and free from irrelevant trends. So, a clean robust dataset improved the performance of machine learning models in downstream workflows and gave better predictions and log reconstructions. Data scientists and geologists have used the interface in Fig. 27 to clean, transform, and QC of the dataset efficiently so that the following analyses were built upon the solid foundation of quality data. This step improves computational efficiency but also decreases the chances of introducing biases or errors into the machine learning models so one can make better decision-making in geological interpretations.

4.3 DETECTION OF OUTLIERS - ANALYSIS

Outlier detection is one of the most crucial steps that comes in the process in developing Log QC and prediction workflow as it detects outliers in data, represented by samples of data that are significantly different from other samples.

Comparing the results of this work, authors have used "One-class Support Vector Machine (SVM)", "Isolation-Forests (IF)", and "Local Outlier Factor (LOF)". LOF is a non-supervised technique, it calculates the difference of the density of a data point compared to its neighbors, it identifies points that have lesser density and provides an anomaly score known as LOF score, and "LOF", unlike other algorithms allows an extent of anomaly to each of the samples instead of marking an outlier or an inlier flag to the sample, that is called Local Outlier Factor (Chowdhury, Tanjil, Akter, Amin, & Pal, 2019). There is one-class SVM or an unsupervised outlier detection method (Popa, Cassidy, & Jikich, 2016) (Ma, 2011) (Eftekhari, 2024); however SVM captures the majority class density it is liable to classify outliers that fall beyond the majority class boundary. In Isolation Forest, IF is also an unsupervised model and like Random Forest it also uses the decision tree based on the fact that abnormal samples are few and distinct. With the help of comparing outputs from three algorithms, the authors realized that results that are most geologically representative and accurate can be provided only by the LOF method due to the fact that such a method doesn't use any kind of prior assumptions about the data distribution and locates local outliers instead of global ones. Comparing the three figures – (LOF), (SVM) and (IF), wherein a total of 14,451 log samples were analysed,

Table 10: Comparison of Total Vs Outlier datapoints through algorithms

Method	LOF	IF	SVM
Total Analyzed Samples	14451	14451	14451
Total Outlier points	195	148	197

4.3.1 IDENTIFYING LOG OUTLIERS USING LOF METHOD

Fig 29: Cross plot of Density Vs Sonic Curve, wherein Outliers have been highlighted in RED color by using LOF method, encircled Graph: Cross plot with "Density" on the vertical axis and "Sonic" on the horizontal axis. This figure 29 is a graphical illustration of the relationship between two basic well log variables that carry important information concerning subsurface geological formations. In this graph, most of the data points have been coloured in green, and

these have formed a tight, dense cluster, which means that the relationship between density and sonic values has strong correlation.

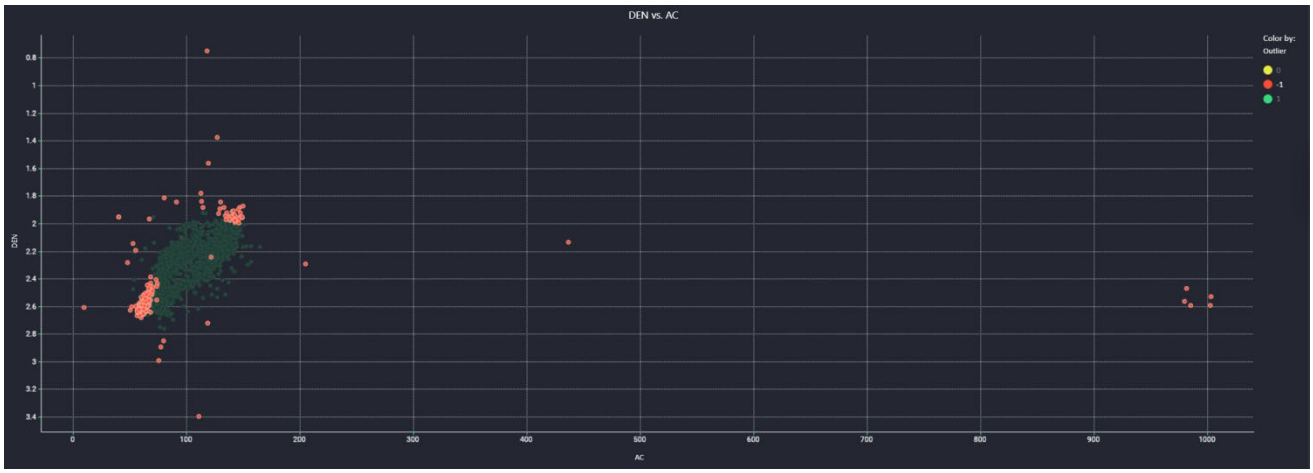


Figure 29: Cross plot of Density Vs Sonic Curve, showing Outliers identified in RED color using LOF method, marked in circles

These green points are the usual geological formations that would tend to follow predictable relationships. In most cases, the subsurface materials like various rock types and fluid follow these trends because of the natural properties of the rocks and the predictable interaction between the sonic and density measurements. Sonic logs measure the speed at which sound waves travel through rock formations while density logs give the bulk density of the subsurface formations. There is a remarkable correlation between the two parameters about their type of geology in terms of porosity, fluid contents, and the other lithology factors in most of the cases. However, as shown in this green-colored cluster, there appear red points which have clearly been marked as outliers. These red points are seen to strongly deviate from the relationship between the variables of density and sonic values.

In this context, outliers refer to points that do not fit the general trends observed elsewhere in the dataset. The identification of these outlier points is very important for several reasons. These red points in here were identified using the Local Outlier Factor algorithm, which is one of the most popular methods in unsupervised machine learning, applied to data with non-uniform distributions for anomaly detection. The LOF method works based on a comparison between the local density of a point and the densities of the nearest points. If the density of a certain point is so much lower than the density of its neighbors, then the LOF algorithm recognizes it as an outlier. In this case, red points have been marked because their local density is much lower in the cross plot than the points surrounding them. Therefore, this method is especially priceless because it assumes nothing about the global distribution of data-it merely

tries to find local violations of otherwise uniform local behaviour. For geological data, often, different formations have different density or sonic relationships-thus, globally outlier scatterings don't say anything useful. The LOF approach provides a more subtle view of the points that actually represent unusual or erroneous values. Such red outliers arise from multiple causes. Measurement errors during data acquisition might be one cause.

Well logging is complex in the sense that it requires one to pass down the instruments along boreholes in a manner that will offer continuously taken measurements. There are many other factors that might introduce errors; it could be equipment failure, environmental conditions, or even error by an operator. Besides these sources of outliers, there may exist unusual geological formations which are not in step with certain trends. For example, there could be zones that are anomalously high in value compared to what is common in more frequent formations like anomalous zones with abnormal compositions of minerals or fractures in certain types of rocks. Fluid anomalies in subsurface conditions may also contribute to the outlier points. Sometimes, outliers may indicate even geological phenomena that are different and rare, such as natural fractures or faults, which need further investigation from geologists with regard to their main reason. Detection of outliers is one of the steps in the data preprocessing workflow since these outliers can significantly alter any further analysis or modeling efforts. If these outliers are not addressed, the models would tend to be skewed and possibly make less accurate predictions and their interpretation of the subsurface geology is likely misleading. For example, in well log prediction or reservoir characterization, using machine learning models, such outliers may introduce unnecessary noises into a model which lowers the accuracy and reliability of the model. An important aspect of data quality concerns the discovery and treatment of outliers-in this case, either by removal or by appropriate techniques for correction. As the robustness of the final models usually depends on such quality in real-world conditions, outliers happen to also be an interesting matter of study: some of these anomalies could reflect important geological features which otherwise have not been so clearly evident, such as fault zones or areas of anomalous porosity.

For instance, once these outliers are studied in greater detail, geologists and data scientists will know what to do-that is, to remove the outliers from the analysis as only noise or pay more attention to them as something of bigger geological interest (Fryklund & Stark, 2020) (Caporin & Fontini, 2017) (Barbosa & de, 2019). In a way, then, the process of outlier detection does not only clean up the data quality but also assures that no valuable feature of geology would have been left behind. In conclusion, this cross plot of density versus sonic is very useful in

distinguishing between green inlier points and red outlier points. This underlines the need to get anomaly information in geological data sets. Again, in the Local Outlier Factor method applied here, the method correctly flags these anomalies as either potential corrections or deeper geological insights depending on the context. Correct handling of outliers is essential, since it allows for appropriate handling in subsequent analysis, modeling, and decision-making tasks related to well logging and reservoir characterization.

4.3.2 DETECTING LOG OUTLIERS USING SVM METHOD

The cross plot with "Density" on vertical axis (labeled as "RHOB [g/cc]") and "Sonic" on horizontal axis (labeled as "DT [us/ft]") is a graphical representation of the association between these two critical log parameters. This figure is vital for detecting outliers applying SVM techniques. That is a complex method in the machine learning family techniques applied in classification and outlier detection. Any data point that departs from the general pattern in the data set might be an indicator of measurement, geological anomalies, or simply other irregularities. This plot indicates that most of the data points fall into the green color and form a dense cluster, representing typical values following the expected relationship between density (RHOB) and sonic travel time (DT). The green points indicate that these data values lie inside the overall trend of observations made on the subsurface. Thus, they must possibly correspond to typical rock formations and their physical properties, considered as typical ones. The RHOB log measures subsurface formation bulk density, whereas DT log depicts the time it takes for sound to travel across the formation. These two logs are highly correlated with each other since denser rock typically has slower sonic travel times. Thus, the green points represent this predictable correlation, and are thus classified as normal data or representative of typical geological conditions under which the figure 30 was drawn.

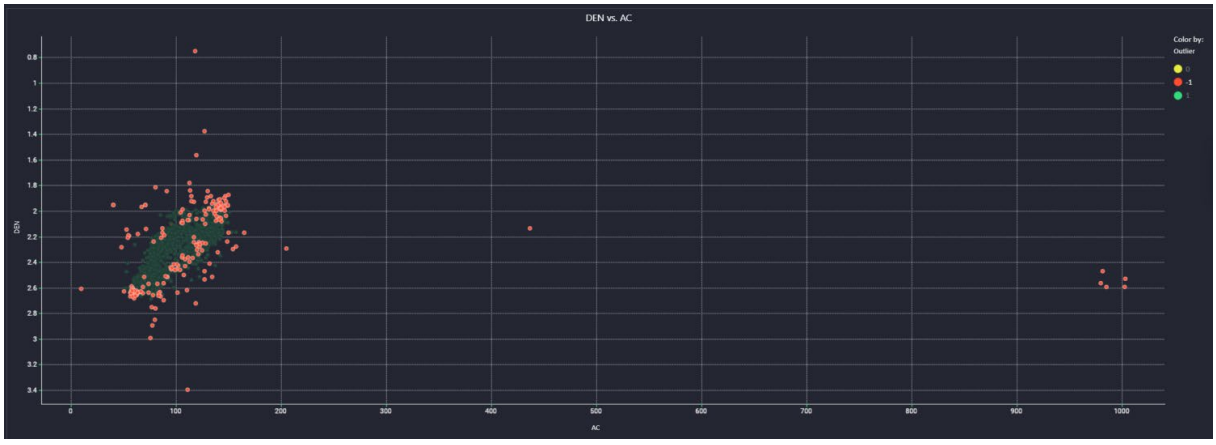


Figure 30: Cross plot of Density Vs Sonic Curve, showing Outliers identified in RED color using SVM method, marked in circles

However, in the midst of all these green spots are red spots which obviously characterize outliers. These red spots are the outliers that were identified by employing SVM classification and are circled in the plot below, illustrating just how strongly they deviate from the bulk of data points. A prime example of this is the SVM method, which is very commonly used in machine learning. SVM is particularly effective at outlier detection since it classifies data points on an optimal boundary—or hyperplane—that separates most of the normal data points from outliers. Here, the red points represent those that the SVM model has picked out as falling outside the boundaries that the majority of the data have defined: These are outlying data points that do not fall in the general trends that can be established between the density and the sonic travel time, and might have major implications to the interpretation of the well log and the underlying geology. The detection of such outliers is important for two major reasons. First, outliers usually point to an issue with data quality, such as errors in measurements or noise during the acquisition process. Well logging is somewhat of a complicated process. Equipment and environmental conditions may fail, or more likely, the human being that is collecting the data might just mess up. These red points could be instances where the logging tools failed to capture the density or sonic travel time measurements with very good resolution and produced values in contrast to the rest of the dataset. If these faulty observations are not identified and corrected, they may interfere with further analysis and interpretation and lead to spurious conclusions about the subsurface formations.

The outlying values could also reflect geological anomalies. For instance, subsurface conditions or types of rocks could have anomalous physical properties such that their data is outliers from the general relationship between density and sonic travel time. This could include heavily fractured zones, gas-bearing reservoirs, or unusual mineral compositions. Such formations may have some peculiar density and sonic characteristics different from the types of rock commonly found in the well, and their identification can provide very useful information about the complexity of the subsurface. In some instances, such anomalies may indicate the presence of economically viable resources, like hydrocarbons or minerals, and hence their identification is quite crucial for both exploration and production. The identification of outliers is also critical to improve the overall quality and accuracy of data. If unchecked, outliers may interfere with the outcome of different statistical analyses, machine learning models, or reservoir simulations. For instance, in machine learning-based log predictions, the

existence of outliers introduces noise in the model, thereby reducing its accuracy and generalizability. Thus, by removing or correcting outliers, geologists and data scientists ensure that their models are built on clean high-quality data, ensuring the reliability of their predictions and subsequent decision processes.

Identifying outliers might, in fact, give useful information beyond reporting what is to be removed. The process of outlier identification can serve as a tool for diagnostics, where it will highlight parts of the dataset that require more intensified examination. For example, the red points may make the geologists pay another look at the parts of the wellbore where anomalies first happened so that they might have a better appreciation of what is happening. For instance, outliers may represent a certain sort of geological feature, such as a fault zone or high-porosity formation, that might have otherwise been overlooked in the exploration. This level of added insight helps geologists make more informed decisions with regards to the subsurface, thus enhancing both the exploration and development phases of a project. In a nutshell, the plot of "Density" vs. "Sonic" with outliers identified by SVM is an effective anomaly detection tool for well log data. Here, most of the normal data are in green points, while the red points represent points deviating from trends. The reason why this is a good approach to achieve accurate data quality and not measurement errors and possible geological anomalies is that the SVM method could possibly be highly efficient in classifying these outliers as an optimal boundary. Therefore, proper handling of these outliers, such as correcting them, removing them, or further investigation, is necessary to ensure that correct analysis and interpretation of subsurface formations are undertaken. Anomaly knowledge will provide to geologists and data scientists better models and better decisions pertaining to the exploration and production processes.

4.3.3 DETECTING LOG OUTLIERS USING IF METHOD

The figure 31 referred to here is a cross plot which has "Density" along the vertical axis and "Sonic" along the horizontal axis, and it is made use of to check for outliers via the Isolation Forest (IF) method. This plot is an important tool for visualizing data points that are markedly different from the expected behavior. Its goal is thus to improve the quality of the data itself while making the further analyses or models more robust. Outliers in well log data are especially significant in subsurface exploration. At times they can represent erroneous data, whereas other times, they may give insight into unique geological conditions. Here the plot points represent pairs of critical density and sonic measurements which are critical logs in

understanding rock properties in the subsurface. Most of the data points follow normal trends as shown by the points in green; however, there are points that are red suggesting outliers which the IF method has identified. In total, there are two points types enclosed in the cross plot: the green points and the red points. The green points are actually the dense cluster inside the plot, which describe the overwhelming majority of the data with a typical predictable relationship between density RHOB and sonic travel time DT. Essentially, density and sonic measurements are correlated in well log data since denser rock formations normally have slower travel times of sonic. So the green points thus represent the expected physical behavior of subsurface. They are normal data that is well usable to build the models further.

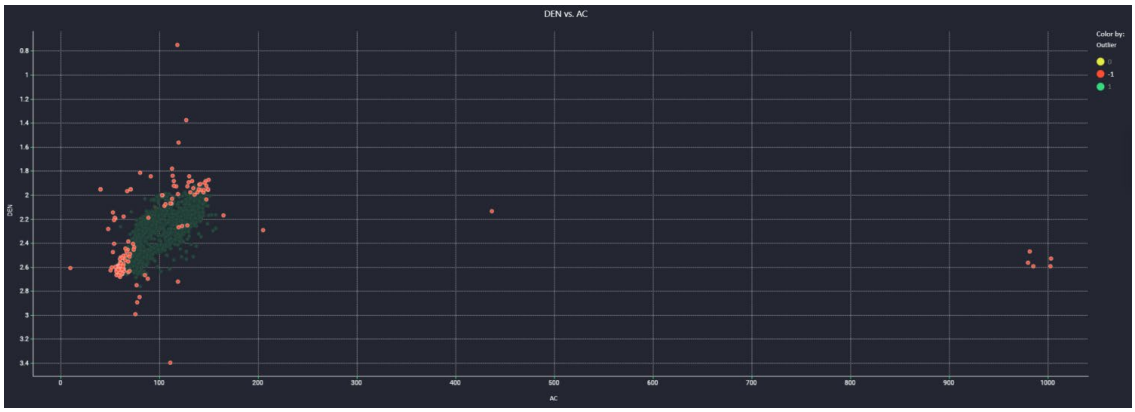


Figure 31: Cross plot of Density Vs Sonic Curve, showing Outliers identified in RED color using IF method, marked in circles

The red points are the outlying ones identified with the help of IF. This red point does not lie close to a cluster of green points indicating that the point does not follow the general trend of the data. The Isolation Forest method is a technique applied to identify the influence that each point has to a particular predictive model. According to the context, in this case, the IF method identifies the points that have an outsized influence on the model's output or performance. In other words, the red points represent data entries that, if removed or altered, would cause a rather large change in the model output. The outliers might be due to various reasons, ranging from measurement errors to just noise in the data and unusual geological formations. Their presence is of the highest importance for several reasons: They point to potential data quality issues first and foremost. Well log data may include some erroneous measurements caused by instrument malfunction, environmental conditions, or errors in the measurement process. For instance, once a logging tool gets faulty during the determination of density or sonic travel time, it will give values that do not agree with the balance of the rest of the data set, and it will appear as outliers on this cross plot. Such anomalies would go undetected and corrected if they

do not surface and affect the quality of subsequent analyses, leading the geologist into a wrong conclusion or incorrect models for subsurface characterization. Other than the measurement error, these outliers may be an actual geological anomaly. Subsurface formations typically do not have uniform characteristics, and sometimes the fractured zones, gas-filled reservoirs, or highly porous formations are unusual in density and sonic values combinations. For example, a gas-filled formation could be less dense and more sonic than the average normal rock formation, giving a somewhat spurious result on the plot. Similarly, some rock types, such as unconsolidated sands or highly fractured zones, may produce anomalous log responses. Identification of such geological anomalies will be crucial for making the accurate model since, through such geological anomalies, a whole area could be lost for resource exploration, well planning, and reservoir management.

IF Method:

The Isolation Forest method for outlier detection is useful because it focuses on identifying points that have a disproportionate impact on the overall model. Traditional outlier detection methods only would identify the points for some distance from the mean or other statistical thresholds, but IF method evaluates on how each point influences model predictions. Thus, finding which data points have most influence on the model, the IF method will ensure the troublesome or impactful outliers be corrected. This is particularly important in machine learning workflows, where the presence of a few influential outliers can cause model performance to degrade dramatically. Once outliers have been identified, several strategies can be followed about what to do with them. For example, outliers may simply be removed from a dataset, provided that their existence is thought to be the result of measurement errors. In other cases, the outlier values may need correction in order to be a good fit for the majority of the remaining data set. However, when the outliers are real geological anomalies, they should be left in the data set but used differently in the analysis. For instance, such outliers could potentially be influential in exploration resources in which unusual formations could be an indication of hydrocarbon or other commercially valuable resource presence.

Data Quality and Modeling Implications

Outlier identification and exclusion is one of the significant phases of data quality that prevent mistakes in measurement as well as geological anomalies from affecting advanced datasets like well logs. As such, improper handling of outliers might cause lots of damage in the resulting machine learning models, analyses, or reservoir simulations. For example, poor data in

reservoir modeling may lead to a wrong estimate of the reservoir size, content of fluids, and hence production capacity, all factors which go into the aspect of feasibility of an economic project. Finally, the cross plot "Density" vs "Sonic" with outliers determined by the IF method could be useful for quality control of data and to indicate possible anomalies in the geological context. The green-colored points indicate normal data, and red points represent significant outliers. The use of the IF method by geologists and data scientists is helpful in developing effective outlier detection, which further impacts the robustness of their models and aids in better decision-making for the selection of the subsurface model that is accurate and for optimization of exploration and production. It makes little difference whether the outliers result from errors of measurement or are genuine geological anomalies-the need for their identification and appropriate handling is felt both for developing an accurate subsurface model and optimizing exploration and production activities.

4.3.4 LOGS OUTLIER DETECTION USING THREE ALGORITHMS

One of the main steps in the current workflow is outlier detection that tries to pin point the outliers from the data mainly those data points that do not follow the general trends and hence look significantly different from the other samples. In the present work, Outlier detection was done through three different algorithms; "One-class Support Vector Machine (SVM)", "Isolation Forests (IF)", and "Local Outlier Factor (LOF)".

Figure 32: Outlier detection workflow with different methods.

The figure 32 referred to as ML QC and Well Selection Workflow depict a sophisticated software interface developed with the objective of optimizing the selection and management of well log data within the framework of the machine learning (ML) workflows. This interface was designed to specifically cater to the oil field, where the interpretations and analyses of well logs are significantly important for subsurface exploration, characterization of the reservoir,

and decision-making in field development. The following figure depicts the workflow for choosing wells and variables for quality control and other stages of ML-assisted tasks, including clustering, outlier detection, and predictive modeling. This interface plays a significant role in streamlining the choice of data, achieving quality control mechanisms, and allows the ML technique to be applied effectively for achieving higher accuracy and reliability in subsurface modeling.

The interface in this figure integrates multiple key components in support of the ML-based workflow. On the left-hand side, users can view a well list stored in the database, which is filtered by specific criteria of "Well Name," "API," "Operator," and "Basin." Filters are applied to select accurate wells; geoscientists and engineers narrow their focus to a subset of wells pertinent to their analysis. Under the filters, a good well list is provided below, showing the details for each well such as its API number, operator of that well, and the basin in which it resides. This section constitutes the beginning point for the selection of wells for further analysis in the interface. A set of tabs at the right end represents different stages of the ML workflow. The current open tab to view is "Well and Variable Selection," where the wells that have been selected for the workflow are shown, with each of them appearing in different colours. One can also look at the specific variable attached to the wells under the section known as "Variable Qualifiers," including such variables as Gamma Ray (GR), Resistivity, Density (RHOB), Sonic (DT), and Porosity (PHI). These are the essential logs that will be used in subsurface evaluation, providing basic information regarding identification of lithology, fluid content estimation, and characterization of the reservoir.

There are other tabs created to support the various stages of the ML process: "Workflow Selection," "Clustering," and "Outlier Testing.". Under the "Workflow Selection" tab, users can choose particular ML workflows that are specifically suitable for selected wells and variables, like regression models for predicting porosity or classification models intended for lithofacies identification. The "Clustering" tab is for unsupervised learning tasks like discovering patterns and groupings in the data. This is useful for detecting geological facies or zones with similar reservoir characteristics. The "Outlier Testing" tab identifies anomalous data points so that erroneous or unrepresentative logs don't contaminate the accuracy of the models being developed. At the bottom of the interface, there is a map area labeled "Map R20k Open." Although this area is currently inactive or empty in this figure, it works to be a helpful tool for geospatial visualization. It would make the map a source of visualization for the spatial distribution of the selected wells; users could then easily attach geographic context to the

selection process. This kind of functionality comes in handy when working with oilfields, in which spatial relationships between wells and geological structures play a large part in carrying out field developments and resource management.

This interface will act as a primary selection and management hub in ensuring the integration of well log data into ML workflows. Efficient management of large datasets, choice of relevant variables, and the use of proper ML models are critical to achieving an accurate subsurface model in oilfield well logging. This interface therefore simplifies the question of selection of specific wells and variables from which only the best quality, relevant data are subsequently used in the workflow. With this structure provided by the interface, geoscientists and engineers will be better placed to optimize machine learning techniques during well log analysis to improve precision in predictions of fluid content and lithology in a reservoir. The importance of this interface can be seen in how it can further improve the overall workflow concerning the management of well log data and the model's development. Data quality and relevance are key factors in the oilfield operation because improper or incomplete data leads to inaccurate prediction and results in more expensive errors. By including filters along with the on-the-run quality control mechanisms, the interface ensures that only accurate and relevant data are selected for input in the analysis. This is even more crucial for ML workflows since their performance strictly depends on the quality of input data.

Several stages included in the workflow at each stage-from well selection to clustering and outlier detection-made the experience a smooth user experience. Experts were allowed to switch between different jobs easily without having to switch between multiple tools or interfaces. This integration allows for a more streamlined and efficient workflow, thus saving time and effort while preparing data for the development of ML models. Moreover, outlier testing and clustering can be done within the same interface enabling the user to identify noise or anomalies in the data that are commonly known to prevail over well log analysis. The interface helps ensure that the ML models are trained on clean data and that they are representative, which improves the accuracy of and reliability of the predictions. Correspondingly, analysis could also make the underlying patterns in the data visible to unveil subsurface heterogeneity and help point out key geological features.

The framework proposed by this interface therefore bears major implications for the oilfield industry. Subsurface model accuracy may significantly be improved by more efficient management of large datasets, rigorous quality control measures, and the application of

machine learning techniques. This, in turn, will better support decision be making on exploration, field development, and production optimization. For instance, accurate predictions of reservoir properties and fluid content permit operators to target the most productive zones, reduce drilling risks, and optimize completion strategies. That is; by streamlining well and variable selection and ensuring data quality, the framework integrates different ML tasks in one interface into an efficient solution for enhancing the accuracy and efficiency of subsurface modeling in the oilfield industry. In conclusion, it represents a sophisticated tool for integration with machine learning in the process of data management of a well log.

4.4 ANALYSIS OF WELLS AND VARIABLES SELECTION

As part of the workflow, wells and log curves were selected for certain areas of the ML workflows – detection of outliers, feature selection, clustering, and ML training for prediction. It was intended to just be a central location managing all of the wells and variables that are available. Some specific wells and variables were chosen for each area of the workflow. Figure:5 Dataset selection for various AI/ML workflow. Dataset available in the database is shown on the left hand of the screen and dataset selected for a particular workflow is shown right hand side of the screen.

Figure 33: ML QC and Well Selection workflow.

The interface "ML QC and Well Selection Workflow" is a new modern oilfield data analysis that will simplify and optimize the selection of wells and log curves for a wide array of machine learning tasks, such as outlier detection, feature selection, clustering, and training models for predictive analysis. This interface not only enhances the efficiency of the whole data preparation process but also ensures that chosen data is of high quality, which is really necessary for developing ML models for a good interpretation of well logs. Below, we describe

in detail how this interface functions, its functionalities, and why the application of such an interface in oilfield industrial structures, as depicted in figure 33, is so important.

- **Available Data in Database**

The left side of the window is a central warehouse for all well data available in the database. It contains several filter options, of which includes "Well Name," "API," "Operator, and "Basin," among which it might apply the needful filters to narrow down the variety of wells to choose from. These filters are very important when managing large datasets; thus, users can easily find and select the wells relevant to their analysis needs. For example, a geoscientist with an interest in wells within a particular basin or which are managed by a particular operator can have these filters aid them in zooming into the list and concentrating on the data that would be most useful for an ML workflow. Under the filter options is a comprehensive list of well names accompanied by other relevant information, like the API number, Operator name, and Basin name. This provides an overview of what data is available. It is this list entirely of wells found in the database on which the wells for use in the ML process will be selected.

- **Dataset Selected for the Process**

To the right of the interface, a set of tabs is provided and used to represent the different phases of the workflow of ML. This ensures that the user advances through data preparation, analysis, and model building in an orderly fashion, ensuring that all the steps of the process are attained. Each tab represents one significant stage of the process and can interactively engage the data and make necessary adjustments at each stage.

Tab 1: Well and Variable Selection – The place where the user makes the initial well selection alongside its associated variable, or log curves. The wells selected from the left part of the interface are coloured differently for easy identification. Under the columns titled 'Variable Qualifiers,' the user will find the exact log variables like Gamma Ray, Resistivity, Density, Sonic, Porosity, etc., corresponding to these wells. This is an important step because it guarantees that only the right log files are brought into the analysis as each variable provides unique information related to subsurface such as lithology, fluid content, and the rock properties.

Tab 2: Workflow Selection – In this tab, users select one or more specific ML workflows they want to apply to selected wells and variables. They might for instance want to apply regression models to predict porosity or lithology classification models to flag types of rocks. It saves

time because the user will be able to choose pre-defined workflows so that they can apply machine learning models to the data, ensuring that the right analytical methods will be applied for the selected well logs.

Tab 3: Clustering- This tab is concerned with the clustering analysis, which remains as an unsupervised type of machine learning technique used in grouping data points having similar characteristics. In well log analysis, clustering may become very handy in defining which zones of the well have similar rock properties; this may be very helpful in defining geological facies or reservoir zones with equivalent characteristics. It is an important step toward understanding distribution of property across the well and to make data-driven decisions on how to model the subsurface.

Tab 4: Outlier Testing This tab detects outliers within the well log data. Outliers can severely distort machine learning models if these are not appropriately identified and handled. Proper outlier testing is therefore essential in ensuring that the data applied during the ML workflow are free from such outliers, reflecting actual subsurface conditions. Detection of and elimination of the outliers can help in improving the quality of the data ensuring higher accuracy from the modeled predictions.

- Map Area:

At the lower portion of the user interface is a map area labeled "Map R20k Open". Although it does not appear populated in the current figure, it is probably used to graphically locate wells analyzed. The oilfield industry uses geographic context as any relationship that exists between the wells and the geological structures will impart some influence on how subsurface data is interpreted. With data populated into this map, it would become a visual powerhouse that would help users understand the spatial distribution of the wells and related variables. This interface serves as a central point that manages the whole ML workflow-from choosing wells and variables up to the use of machine learning techniques, including outlier detection, feature selection, clustering, and predictive modeling. The use of this interface ought to lead to increased efficiency and accuracy in handling the data preparation work so that high-quality, relevant data are selected properly for analysis. This is quite significant in the oilfield since there will usually be huge datasets, and the quality of the data directly translates to the quality of the ML models.

This interface is very fundamental to the oilfield because it involves the interpretation of accurate interpretation of well log data for reservoir characterization, exploration, and field

development. To improve the accuracy of subsurface models, the following well management and selection criteria and also rigorous quality control mechanisms should be applied. It should easily integrate the various stages of the ML workflow. Using its well-structured tabs, users can navigate through the different stages of their workflow easily and hence gain productivity.

This interface also enables geoscientists and engineers to deduce interpretable insights from well log data because it allows users to use ML workflows including clustering and anomaly detection. For instance, through the use of clustering analysis, a geoscientist may be able to identify patterns within the data which might not be quite apparent, such as zones having similar properties or facies in the rocks. Outlier detection ensures that the training data used in applying ML models is clean, since the erroneous points might distort predictions. Conclusion The interface "ML QC and Well Selection Workflow" is a powerful tool for managing well log data and application of machine learning techniques in the oilfield industry. Improvements in subsurface modeling efficiency and accuracy arise through this interface: it streamlines the selection of wells and variables, integrates the various stages of the ML workflow, conducts outlier testing and clustering for improving data quality. The potential to process large datasets and apply direct sophisticated ML techniques into well-log interpretation makes this a very valuable tool for geoscientists and engineers engaged in exploration and development work on an oilfield.

4.5 FEATURES SELECTION ANALYSIS

Features Selection: another essential step in the existing workflow, in which the predictive model is built without any correlated variables, biases, and unwanted noise. Within the features selection steps, five different methods were applied: namely Boruta, Correlation Coefficient, Lasso, Principal Component Analysis (PCA), and Variance Threshold, and the results were compared for use in Clustering (Zhou, 2020) (Tabasi, 2022) (Sebtosheikh, Motafakkerfard, Riahi, Moradi, & Sabety, 2015).

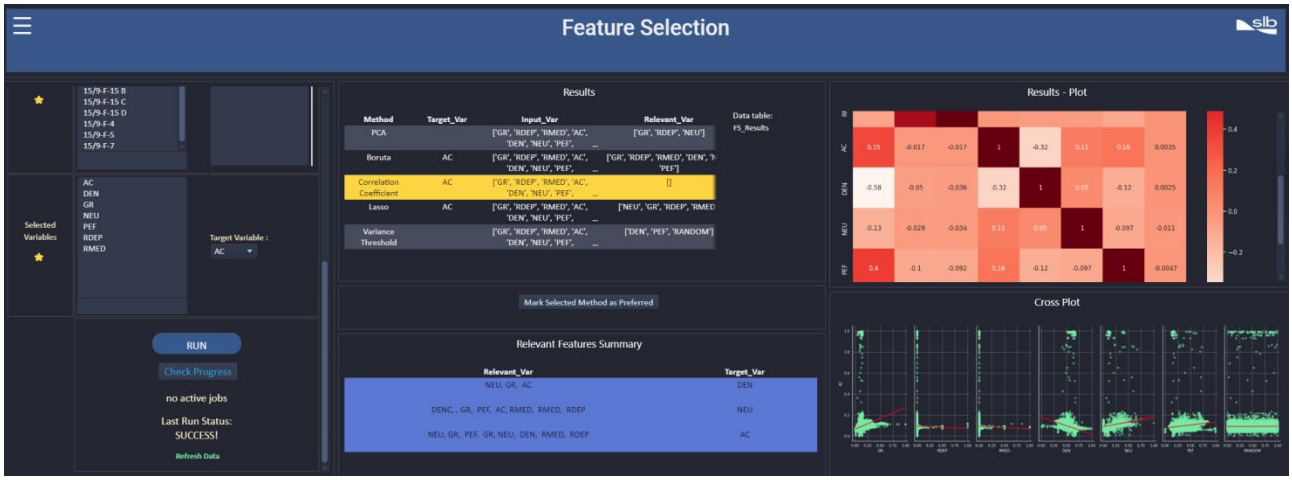


Figure 34: Feature Selection workflow comparing results from various methods highlighting the key results after the Feature Selection workflow was run to compare results from all 5 algorithms

The figure 34 below gives a structured view of the results produced by applying five different feature selection methods to one dataset: Boruta, Correlation Coefficient, Lasso, PCA, and Variance Threshold. Each has a different purpose: to select more features relevant than others for any predictive modeling task, to reduce noise, to eliminate biases, and to train the model with the most important variables. This is of paramount importance in improving the accuracy and efficiency with which machine learning models can be applied to deal with redundant or irrelevant features found in such large and complex datasets as those obtained in oilfield well logging (Rostamian, Heidaryan, & Ostadhassan, 2022) (Otchere D. A., Ganat, Ojero, Tackie-Otoo, & Taki, 2022) (Misra, Osogba, & Powers, 2019). Below is a detailed explanation of the results, the methodologies applied and their overall significance.

1. Boruta: Boruta is the all-relevant approach to feature selection where it can compare the importance of the features with so-called original features and those generated at random, labeled as shadow features. The most poorly performing features on each round are removed from the algorithm since only clearly important features are left. Here you will find selected features represented as Boruta, so you will learn which variables were considered important for the predictive model. Boruta is efficient since all the potentially relevant features are kept without dropping any feature that is not dropped unnecessarily.

2. Correlation Coefficient: Feature selection by correlation coefficient involves a criterion for selecting based on the statistical association of the features with the target variable. Thus, such features that correlate highly with the target variable are considered more predictive and thus selected. Low correlation and high multicollinearity features characterised by high correlation

with other features might be dropped to avoid redundant information skewing the model. The results of this method ensure that the model uses only independent, non-redundant variables with a high correlation to the predicted outcomes.

3. Lasso (Least Absolute Shrinkage and Selection Operator): Lasso is one of the strongest regularization techniques. It uses the L1 norm, shrinking the coefficients for less important features to be zero. That selection of features is therefore by-product of trying to reduce the overall model to the least possible number of features, choosing some based on the negligible impact they have on the accuracy of those modeled predictions. The Lasso method is particularly valuable when there are many features because it can do both feature selection and regularization all at once. There's less chance of overfitting, and the model improves in interpretability. In this pipeline, it displays chosen features and shows which ones have non-zero coefficients and are selected.

4. Principal Component Analysis (PCA): Principal component analysis is a type of dimensionality reduction in which the original dataset is projected onto a smaller set of orthogonal directions often called principal components. These principal components are directions of maximum variance in the data. Most information of the data can thus be preserved by using PCA to reduce its dimensionality. In this workflow, the top scoring principal components are selected for further analysis. Although PCA does not select the original features explicitly, it only provides a compressed version of the data that would yield the essential patterns that are crucial for driving clustering and other predictive tasks.

5. Variance Threshold: The method eliminates features that have very low variance as well, for the assumption would be made that these features do not hold much useful information to perform the prediction. A feature with low variance would imply values of that variable do not vary across much of the dataset; hence, less likely to add value to the strength of the model as regards predictive power. Feasible features going over this threshold will be retained and included in the list within the workflow.

The heatmap at the top center of the interface gives an indication of correlation between different features. Darker shades of colouring on the heatmap indicate a higher correlation between variables, be it positively or negatively. Such a visualization would help users easily figure out which features are highly correlated with each other and thus redundant or should be discarded to prevent multicollinearity in the model. In predictive modeling, multicollinearity

can distort the real relationship between variables and target therefore recognition and handling of related variables is very crucial.

Most Selected Features: Summary of most frequently chosen features across all five methods. This view of the interface summarizes the most frequently chosen features across all features of all the five methods. By comparing the output of multiple feature selection techniques, it depicts which features are selected by several methods. This set of relevant variables is all that has been included for the task of prediction. Other features may be either removed or chosen by a single method that are less informative or altogether irrelevant. This summary, therefore, ensures that the final model contains only the most useful features to perform better and be more meaningful. The bottom right corner of the figure likely represents the distribution or importance scores of features across the feature selection methods. The histograms thus allow the user to visually compare the importance of each feature as determined by the various methods, offering extra insight into how different selection methods weigh different features. This visualization helps to clarify which features are always ranked important between methods, thus aiding final decision making in feature selection.

"The Feature Selection Workflow is intended to improve the quality of the dataset upon which you want to apply machine learning by removing the set of redundant, irrelevant, or noisy features (Handhali, Aamri, & Sundararajan, 2023) (Farsi, 2021) (Dunham, Malcolm, & Welford, 2020). It will run multiple feature selection techniques, each of which might give a different view of what it means for a feature to be important.". By comparing the results of these methods, the workflow thus ensures that only significant features relevant for further analysis are picked. A large number of datasets get generated in applications like oilfield well logging; thus, removal of irrelevant features makes models more accurate and efficient in terms of machine learning. The significance of the workflow is in its wholesome and multi-directional analysis of feature importance. It applies five different methods such as Boruta, Correlation Coefficient, Lasso, PCA, and Variance Threshold so as not to fall into some biases due to the choice of a single method (Bagheri & Riahi, 2015) (Arkalgud, McDonald, & Brackenridge, 2021) (Matinkia, Hashami, Mehrad, Hajsaeedi, & Velayati, 2023). For example, Lasso would preferentially retain fewer features due to the regularizing mechanism, whereas Boruta would prefer to retain most of the features because of their relative importance. The combination of these two techniques is what ensures that one ends up with only the most important features and reduces correlations, noise, and biases in variables.

This further improves the efficiency of cluster algorithms and other machine learning algorithms since it can focus directly on only the most informative features, thus faster computation time, more interpretable models, and accurate predictions. It is especially fundamental in fields like oilfield well logging, where predicting subsurface properties correctly is a necessity. To sum up, feature selection workflow is one of the critical steps in a machine learning pipeline while building models based on solid foundations of relevant and highly quality data (Rahaman, Vasant, Jufar, & Watada, 2020) (Nyakilla, 2024) (Miah, Ahmed, Zendehboudi, & Butt, 2020).

Feature Selection

The process for feature selection in terms of choosing relevant features corresponding to input in ML algorithms is called "Feature selection". The objective of the feature selection in ML is to decrease the number of input features by removing unwanted or irrelevant features thereby making the ML models easy and increasing the time complexity involved with the training ML, thereby increasing the precision of these models. In the following work, authors used and applied 05 different approaches - Boruta, Correlation Coefficient, Lasso, Principal Component Analysis (PCA) and Variance Threshold to assist with the selection of the most suitable and relevant features of the clustering and log prediction stages (Hinkka, Lehto, Heljanko, & Jung, 2018) (Dorrington & Link, 2004) (Anifowose, Adeniye, Abdulraheem, & Al-Shuhail, 2016).

Method	Target_Var	Input_Var	Relevant_Var	Preferred_Flag
Boruta	DEN	['GR', 'AC', 'DEN', 'NEU', 'RANDOM']	['GR', 'AC', 'NEU', 'RANDOM']	
Correlation Coefficient	DEN	['GR', 'AC', 'DEN', 'NEU', 'RANDOM']	[]	
Lasso	DEN	['GR', 'AC', 'DEN', 'NEU', 'RANDOM']	['GR', 'AC']	
Boruta	AC	['GR', 'RDEP', 'RMED', 'AC', 'DEN', 'NEU', 'PEF', ...	['GR', 'RDEP', 'RMED', 'DEN', 'NEU', 'PEF']	Yes
Correlation Coefficient	AC	['GR', 'RDEP', 'RMED', 'AC', 'DEN', 'NEU', 'PEF', ...	[]	
Lasso	AC	['GR', 'RDEP', 'RMED', 'AC', 'DEN', 'NEU', 'PEF', ...	['NEU', 'GR', 'RDEP', 'RMED']	
PCA		['GR', 'RDEP', 'RMED', 'AC', 'DEN', 'NEU', 'PEF', ...	['GR', 'RDEP', 'NEU']	
Boruta	NEU	['GR', 'RDEP', 'RMED', 'AC', 'DEN', 'NEU', 'PEF', ...	['GR', 'RDEP', 'RMED', 'AC', 'DEN', 'PEF', 'RANDOM']	
Correlation Coefficient	NEU	['GR', 'RDEP', 'RMED', 'AC', 'DEN', 'NEU', 'PEF', ...	[]	
Lasso	NEU	['GR', 'RDEP', 'RMED', 'AC', 'DEN', 'NEU', 'PEF', ...	[]	
Variance Threshold		['GR', 'RDEP', 'RMED', 'AC', 'DEN', 'NEU', 'PEF', ...	['DEN', 'PEF', 'RANDOM']	

Figure 35: Feature Selection process results after successful execution and Output was compared in the form of relevant_var.

Feature selection process in machine learning (ML) is an important step to enhance performance as it identifies the most relevant features to retain and eliminate the redundant or irrelevant ones. This is an important step, particularly for applications such as oilfield well logging that contains a large number of variables within the datasets-precisely many of which may not add meaningfully to the prediction tasks. In your work, you applied five different methodologies-Boruta, Correlation Coefficient, Lasso, Principal Component Analysis (PCA), and Variance Threshold-to eliminate more significant features of the data set. This is intended to minimize complexity, increase interpretability of models, speed up runtime during training, and hence eventually increase the accuracy of making predictions.

Objective of Feature Selection: Feature selection is one of the vital stages in constructing a model in ML. The main objective of this stage is to minimize the number of input variables or features of the data while retaining those most predictive for the target outcome. Such reduction in data dimensionality makes the models simpler as well as less prone to overfitting, thus improving generalizability over unseen data. The reduced models also require lesser computational power; hence the training process gets faster along with more efficient prediction. In your case, the feature selection process aimed to enhance the predictability of certain important oilfield well logging variables such as Sonic (AC), Density (DEN), and Porosity (NEU) for the models.

Methodologies Used

1. Boruta: Boruta is one of the more robust all-relevant feature selection methods in which features are selected based on their importance over randomly generated shadow features. In this respect, features that do not perform any better than random shadow features are eliminated, while features that demonstrate significant predictive power are retained. The advantage with Boruta is that this is a conservative method of feature selection such that no relevant features are discarded too early in the modeling process.

2. Correlation Coefficient: It chooses the feature using the statistical relationship between a feature and the target variable. Those features that feature a strong positive or negative correlation with the target variable will be retained, and those shows weak or no correlation will be removed. This method is very effective in identifying variables that show a direct relationship with the outcome of interest, such as the correlation between Sonic (AC) and other variables like Density (DEN) and Porosity (NEU). This method reduces features whose value

is weakly correlated with the target such that the model becomes simpler without losing its predictive accuracy.

3. Lasso: Lasso is an abbreviation for Least Absolute Shrinkage and Selection Operator. It is a technique of regression analysis where L1 regularization applies to shrink the coefficients of any less important features to zero. In this way, the method automatically removes the irrelevant variables while retaining the significant ones that contribute much towards the performance of the model. Lasso is a highly efficient algorithm to deal with high-dimensional datasets, since this mainly reduces overfitting by the reduction of the number of features on inclusion in the final model. Applying Lasso in this workflow reduced some of the variables but kept those variables that had greater influence on the predictions related to target variables.

4. PCA is a dimensionality reduction technique that aims at transforming the original features to a set of orthogonal principal components, where each new attribute will capture some other different aspect of the variance of the data. It is particularly useful for big datasets that contain highly correlated variables, as it reduces the dimensionality of the data significantly with minimal loss of information. Where PCA does not preserve the original features, it does provide a compressed representation of the dataset that may, however still be used for clustering and prediction tasks.

5. Feature Variance Threshold: This technique removes the features with low variance on the assumption that such features would give useless information. Features that have little to no variation in them are unlikely to help in making the process of predicting the target variable effective since they fail to delineate a case from another. Therefore, this technique brings about even lesser dataset dimensions since all those features that carry the least quantity of information will be completely eliminated.

Important Target Variables: Three essential target variables were used in the feature selection: Sonic (AC), Density (DEN), and Porosity (NEU). These are mainly necessary for understanding the properties of subsurface oilfield exploration. With the five methods of feature selection, the idea was aimed at narrowing down relevant predictors for these variables in order to have much-accurate and reliable predictions with resulting models.

In Figure 35 Feature Selection Process Results the results of the process are presented. Figure 35 shows which features have been selected by each of the five methods, based on comparison

of their importance considerations. It cross-references Boruta, Correlation Coefficient, Lasso, PCA, and Variance Threshold results to a very well-rounded view of the most crucial features for further analysis in terms of clustering and log prediction. Correlation and trends of the target variable Sonic (AC) with other features, such as Density (DEN), Porosity (NEU), and the Photoelectric Factor (PEF), are emphasized on. These plots give a intuitive feel for the interactions between the input features with the target variables, so that one can obtain a view of which inputs tell the most about the targets. Figure 13, for instance, shows that Sonic (AC) is very highly correlated with both Density and Porosity, meaning that these are the most significant inputs to apply in the modeling.

Feature selection in this kind of framework has many worthwhile side benefits:

- **Better Model Performance**: It becomes more efficient and accurate for the models, because only the most relevant features are selected. Decrease in noise and variables irrelevant lead to better generalization and improved performance on new data.
- **Reduced Complexity**: The presence of many features which are redundant lowers the complexity of the models thereby decreasing the associated computational costs with training and training times. This is especially when working with big datasets used in oilfield well logging.
- **Better insights**: The knowledge acquired on how the variables interact is also assisting in the process of modeling, as it makes better informed decisions. Features selected improve model accuracy and also help in understanding which variable-driving factors apply to which target variables, just like how a strong correlation exists of Sonic and Density variables.

This proposed feature selection framework will, therefore, ensure that only the most relevant and significant features are applied in the ML model. This, therefore, enhances both efficiency and accuracy of the models, giving reliable predictions and meaningful insights into subsurface properties. Thus, this process remains one key step forward toward further improvements in the performance of machine learning tasks in oilfield well logging and other data-intensive domains.

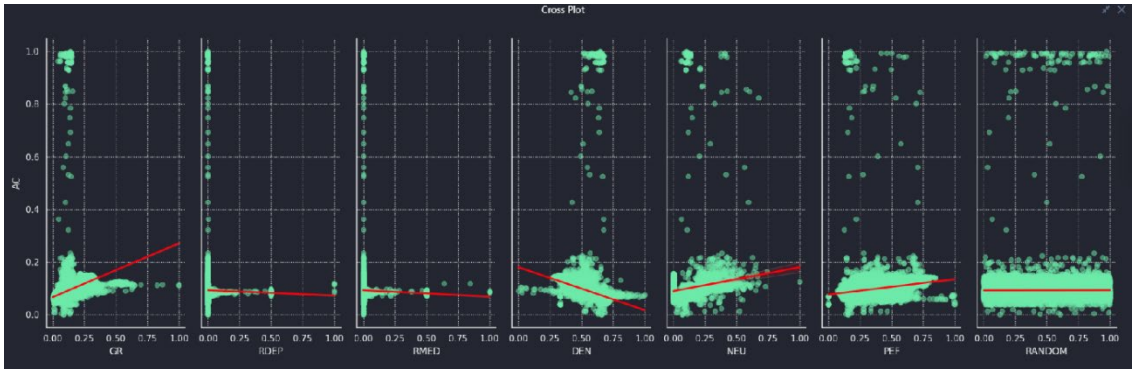


Figure 36: Cross Plot of Sonic Curve (AC) along y-axis Vs Various input curves (x-axis) along with predicted trend line

Figure 36. "Cross Plot of Sonic Curve (AC) along y-axis Vs Various input curves (x-axis) along with predicted trend line" gives a well-rounded view of the relations between the Sonic Curve (AC) and several other well log curves, that is, Density (DEN), Gamma Ray (GR), Neutron Porosity (NEU), Bulk Density (RHOZ), Resistivity (RT), and Resistivity Ratio (RXO/RXA). These cross plots are more important for understanding how various well log parameters interact with Sonic (AC), and this cross-plot can provide valuable insights for predictive modeling in oil field well logging. Break Up:

Each of the six scatter plots in the figure is meant to demonstrate how one of the well log variables on the x-axis might relate to the Sonic Curve (AC) on the y-axis. Visual aides assisting in summarizing the general pattern or trend in relationships presented between variables are represented by the red trend lines drawn through data points within each plot.

1. Sonic (AC) vs. Density (DEN):

One of the fundamental well log curves used to estimate the physical properties of subsurface formations: One can identify one of the basic well log curves in the first plot of the relationship between Sonic (AC) and Density (DEN). Scattered data points set apart along a real measurement plot, but there is a predictive relationship established by the trend line of interest between the two variables. In general, Sonic (AC) and Density (DEN) are inversely related, such that with an increase in density, the travel time for sonic decreases, which means that the speed is higher through denser material. This becomes significant while deciding on rock types as well as mechanical properties.

2. Sonic (AC) vs. Gamma Ray (GR): The following plot discusses the relationship of Sonic (AC) with Gamma Ray (GR), which indicates the natural radioactivity of the formation. The scatter of points and the fitted trend line illustrate how changes in gamma ray readings-which typically corresponds to shale content-may impact sonic travel time. The predictive trend is important because high values of gamma rays normally correlate with slower sonic velocities by higher clay rich layers which has reduced velocity as opposed to clean sands or carbonates.

3. Sonic (AC) vs. Neutron Porosity (NEU): The third plot indicates the relationship between Sonic (AC) and Neutron Porosity (NEU). Neutron porosity measures the content of hydrogen in the formation, often associated with pores filed with liquid. The trend line indicates by how much porosity and sonic velocity are correlated. For all practical purposes, higher porosity may correspond to lower sonic velocities. This makes intuitive sense because the more porous the formation is-in other words, the more fluids or gas it contains-the slower acoustic waves propagate through it than through compact, lower-porosity rocks. Understanding this relationship is important for reservoir characterization and the prediction of rock properties.

4. Sonic (AC) vs. Bulk Density (RHOZ): The fourth plot analyses the relationship between Sonic (AC) and Bulk Density (RHOZ). Bulk Density is the overall density of the formation, comprised of both solid matrix and fluid-filled pores. Like the first plot, this plot probably follows an inversely proportional relationship where higher bulk density lowers the sonic travel time, with denser formations transmitting the sonic wave more quickly. This relationship comes in handy for understanding the rock fabric, lithology identification, and inference of the reservoir quality.

5. Sonic (AC) vs. Resistivity (RT): This fifth plot, the Sonic (AC) vs. Resistivity (RT), is related to the relationship of Sonic (AC) with Resistivity (RT). Resistivity is for the fluid identification in pore spaces, normally the high resistivity values representing hydrocarbons and low values representing water-filled formations. The trend line here explains how resistivity differences influence the sonic travel time. In general, hydrocarbon saturated zones tend to have higher sonic velocity than water saturated zones because of fluid property differences. This crossplot can be used to distinguish between hydrocarbon saturated zones and water-saturated zones.

6. Sonic (AC) v/s Resistivity Ratio (RXO/RXA): The sixth and final plot illustrates the relationship between Sonic (AC) with Resistivity Ratio, RXO/RXA, that is, the comparison between the resistivity of the flushed zone, RXO with the uninvaded zone, RXA. The trend line aids in identifying the relationship between sonic travel time and these resistivity values,

that may imply invasion or fluid displacement profiles in the formation. The resistivity ratio can point to the presence of hydrocarbon, fluid flow, or any invasion of mud filtrate in the formation while drilling. Correlation analysis and predictive modeling are possible by using these cross plots given in the figure. Plotting Sonic (AC) against various input curves makes it easier to visually assess how the variables are related to one another.

These fitted trend lines will form a quick reference point in understanding whether the relationship between them is positive, negative, or weak and hence is vital in determining which of the variables would be most significant for prediction of Sonic (AC). This is useful in preparing data for machine learning models, allowing informed feature selection. Such plots aside from understanding the relationship between variables help pinpoint trends that may go unnoticed in raw numerical data. It is such visual presentation that provides geoscientists and engineers with more control over what to think of the quality of their data, potential outliers, and the general trend of the formation. 1. Correlation Analysis:

Cross plots reveal insights into which of the inputs have the strongest correlation with Sonic (AC). This would be useful in constructing predictive models of oilfield well logging. For example, knowing there is a very strong negative correlation between Density (DEN) and Sonic (AC) would inform feature selection for models designed to predict rock properties or formation characteristics. 2. Predictive Modeling: Trend lines fitted to the data points in each plot then act as an initial indication of the predictive relationships between the input variables and Sonic (AC). They are instrumental components in developing accurate models for machine learning since they aid in establishing how alterations in one variable can affect the target variable, Sonic (AC). These crossplots are crucial in enhancing the precision and dependability of predictions made by subsurface modeling. 3. Data Interpretation:

Such crossplots also make the interpretation of complex well log data more concise by giving a tool for visualization to evaluate relationships and trends. Such visualization helps the geoscience professionals, petrophysical professionals, as well as reservoir engineers, to gain insightful knowledge of the subsurface, which, in turn allows informed choices of exploration and production. Hence, cross plots illustrated in this figure are detailed representations of the relationship between Sonic Curve (AC) and many other well log curves. Scattered data points fit with trend lines in such plots reveal critical predictive relationships that could constitute a needful pre-requisite to construct robust models for oilfield well logging by machine learning.

This analysis helps in feature selection and understanding of correlation, and therefore allows for more accurate predictions with enhanced decision-making in subsurface exploration.

4.5.1 SELECTION OF FEATURES USING CORRELATION COEFFICIENT

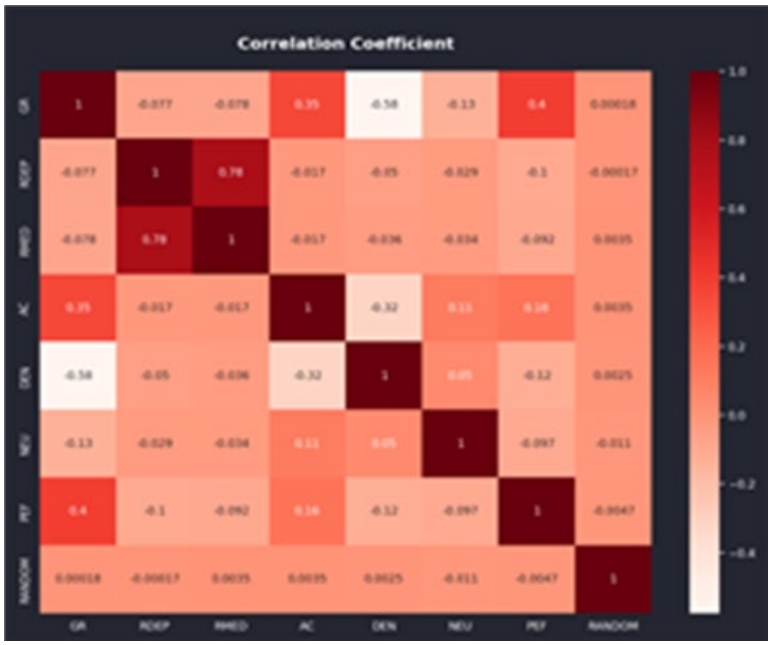


Figure 37: Results from Corr-Coeff, for Feature Selection

Fig. 37 Correlation coefficient matrix Summarizes relationships of various input variables in the context of ML workflows: A critical tool for selection to help recognize features that may be very important to predictors in predictive modeling, especially for applications like oilfield well logging.

Correlation Coefficient Matrix

The goal of the correlation coefficient matrix is to quantify the direction and strength of linear relationships between pairs of input variables. Each cell in this matrix corresponds to one unique pair of variables; the value within each cell represents the correlation coefficient, ranging between -1 and 1: - Positive Perfect Correlation (1): The value of 1 represents a perfect positive correlation between two variables. In the practical world, this results in increasing proportionally as one variable increases. For instance, a strong positive correlation between sonic velocity and density may imply that higher density formations tend to have faster sonic wave propagation. A perfect negative correlation scores -1, indicating that as one value is increasingly higher, another's value is increasing in a consistent, negative manner. For instance, porosity and sonic velocity: perhaps greater porosity occurs in the presence of slower sonic travel times because there is more fluid in the pores.

- Zero Correlation (0): The value of 0 signifies there is no linear correlation between the variables. That means changes in one variable do not systematically relate to changes in the other.
- The matrix also uses a color scale to represent the strength of the correlations in direction, making the data easier to interpret:
- Dark Red: When dark red shows high positive correlation, it means variables moving together are easy to notice.
- White: Indicates a lack of any relationship. This assists in the identification of pairs of variables that are independent from each other.
- Dark Brown: This informs for a highly negative relationship and notifies the use of pairs of variables that have an inverse relation with each other.

The colour-coding is helpful in indicating a potential presence of significant relationships quickly, which should be further looked at or maybe even call for feature deletion in the modeling context. The first and foremost use of the correlation coefficient matrix is to aid with feature selection, that is, to determine which of the variables are most closely related to each other. This is very important for several reasons:

1. Deal with multicollinearity: If the regression model includes two independent variables which are highly correlated, then a high multicollinearity results. It produces inaccurate coefficient values and does not facilitate an easy interpretation of the model. Features with high correlation can be identified using a correlation coefficient matrix. Some of them can be removed; others can be combined. In this way, the stability and interpretability of the model increase.

2. Model performance: The selection of features most relevant will have the effect of improving the accuracy and efficiency of predictive models. The features that are not highly correlated with the target variable would be noise and redundant and will help in degrading the model. To improve the predictive power of the practitioner's models, features that have the strong relationship with the target variable can be emphasized.

3. Model Simplification A more straightforward model with fewer input features is much easier to train, interpret and maintain. The correlation coefficient matrix gives the information that

could be needed in the process of streamlining the model-building process by eliminating features that are unnecessary as such improved generalizability of the model to unseen data.

Importance of the Correlation Coefficient Matrix:

1**. Feature Selection:** The matrix is helpful in feature selection while applying it to clustering and log prediction purposes. Practitioners can take informed decisions regarding what variables to consider and exclude from their models based on relevant features and their interrelations that eventually lead to better outcomes.

2. Interpretation of results: From the correlation coefficient matrix, critical insights about the relationship between variables can be derived. It therefore builds up for more profound insight into data. For instance, when good correlations among some of the input curves are found with the target variable, say Sonic, it is used to steer geophysical interpretation and points out significant parameters affecting reservoir characteristics.

3. **Improved Decision Making**: Clarity of relationship between variables can give clearer insights to managers regarding the implications of these relationships for exploration and production strategies. Implications of these relations for exploration and production can lead to more informed strategic and decision-making related to reservoir management, drilling operations, and resource extraction.

In conclusion, the correlation coefficient matrix of Figure 3 offers an expression for a strong analytical tool within the feature selection process for use in machine learning workflows.

Input variable relationships are quantified by the matrix, and give data scientists and geoscientists the ability to work toward discovering multicollinearity, improving model fit, and simplifying models beyond which predictive power is lost. Colour encoding also offers a further interpretive layer, so practitioners may quickly find higher-order relationships and make corresponding feature inclusion decisions. Finally, the knowledge obtained from this matrix leads to greater data understanding and increased predictive capabilities in oilfield well logging applications.

4.5.2 SELECTION OF FEATURES USING PCA

Figure 38 illustrates a bar chart that describes the amount of variance each PC captures in a PCA. It is one of the most powerful dimensionality reduction techniques used to simplify

complicated datasets and preserve their most important information. The information presented in the bar chart aids in understanding the amount of variance in the original dataset that can be captured by how much of each principal component, hence gaining insight into the underlying structure of the data.

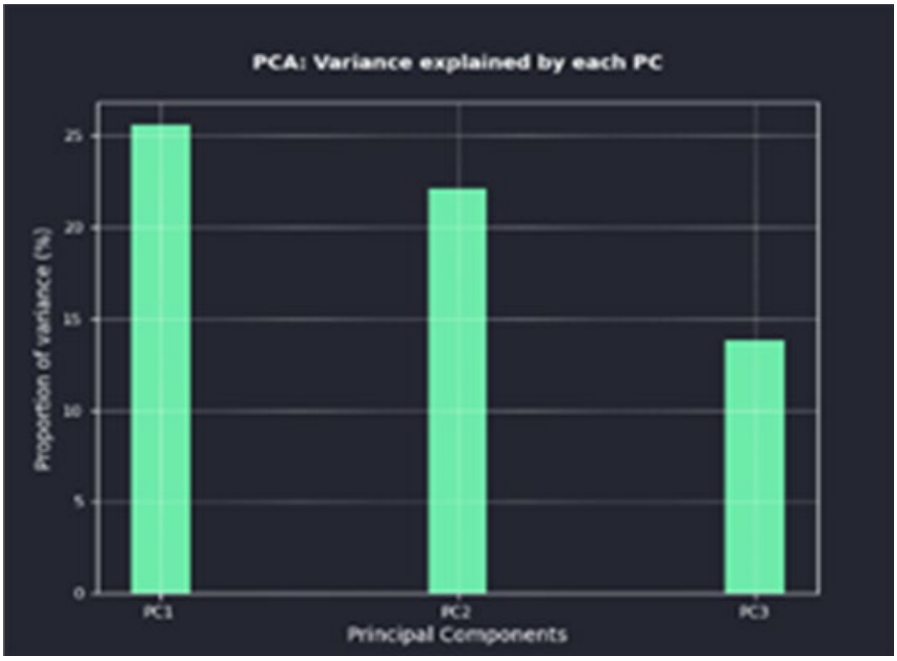


Figure 38: Results from PCA for Feature Selection

- **Y-Axis (Percentage of Variance [%])**

The Y-Axis for the bar chart represents the percentage of total variance explained by each principal component. The graph scale varies from 0% to 25%, thereby indicating the components explain a considerable amount of the overall variability in data. Since this is a significant axis, it would describe the importance of each component and the new features it conveys for the data set.

- **X-Axis (Principal Components)**

The X-axis has listed the principal components in order, as PC1, PC2 and PC3.Each bar represents a principal component with its related variance contribution. The plot can then easily compare the explanatory power of the components.

- **Principal Components**

PC1:

The first bar of PC1 explains variance slightly above 20%. From this, one can interpret that PC1 is responsible for the variation describing most of the features. For most datasets, the first principal component often tends to embrace the patterns of variance, hence a basis for most other analyses.

- PC2:

The second bar for PC2 describes just below 20% of the variance, accounting for the next most significant variation after PC1. It contributes to less variance than the PC1 but its role is still massive and offers more insight into the structure of the dataset.

- PC3:

Finally, the third bar is that of PC3, which accounts for about 10% of the variance. PC3 takes in less variation than either of its predecessors, but the information it does contain is worth enough to be able to reveal patterns or relationships underlying data that are not apparent with the original variables.

The primary aim of PCA in feature selection and dimension reduction is to change the dataset into a smaller set of principal components, which are mutually uncorrelated and possess most of the original data variance. Transformation thus carries out a lot of important functions:

1. Simplify the Model:

It reduces the number of features thus enabling better model interpretation and reduced overfitting. Simplified models usually require less computational power and also can be trained more effectively.

2. Performance Improvement:

This focusing of attention to the major components will also reduce noise, thus decreasing multicollinearity that may complicate the interpretation of feature contributions in the model. A greater proportion of explaining variance by the emphasis on the components can increase the model's predictive accuracy.

3. Visualization of the Data:

PCA proves to be beneficial in projecting the data onto a low-dimensional space that can be applied significantly in exploratory data analysis. There, the analyst may find it easier to

describe the underlying structure using the patterns, clusters, and relationships that were not visible in higher dimensions.

Importance of PCA in Feature Selection

1. Feature Selection:

PCA facilitates the determination of the most important features, which describe most of the data variance. Practitioners may reduce their dataset by considering only those principal components that account for the majority of the variance, thereby developing better and more efficient predictive models.

2. Data Interpretation:

It enables understanding how much variance each principal component explains in order to be able to interpret the hidden structure of the data. For example, if some variables are contributed predominantly to the first principal component, that will help in further analyses and tasks like feature engineering.

3. Understanding Feature Relationships:

PCA not only reduces dimensions but also explains how the features are related to each other. Once data scientists inspect what sort of contribution the original feature provides to the principal component, they will find which features will be influential and how they interact with each other and thus contribute towards better decision-making in modeling.

Base on the information provided in the above section, the strength of every principal component to explain the variance makes the bar graph of the result of PCA stand out in feature selection. It thus is a basis for analyzing complex data sets. PCA, therefore, is a technique for dimensionality reduction focusing on significant variations by making the modeling process straightforward and performance-enhancing and also helps in visualizing data. The insights gained from this analysis are invaluable in feature selection, offering a pathway towards the building of strong predictive models with interpretability retained. It is one of the more important elements of the proposed framework in the context of oilfield well logging and applications. This way, more intelligent decisions can be taken as PCA results in better performance in both exploratory and predictive analytics.

4.5.3 SELECTION OF FEATURES USING LASSO AND VARIANCE THRESHOLD

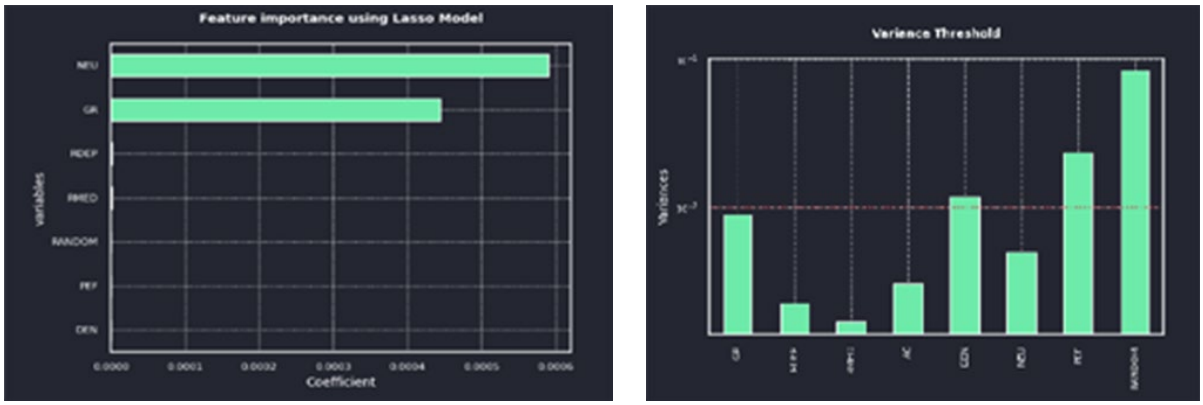


Figure 39: Results from Lasso and Variance threshold for Feature Selection

Figure 39 graphs a side-by-side comparison of feature importance as computed via two different techniques: the Lasso regression model and the Variance Threshold. Two techniques are applied here in the step of feature selection, which is an essential component of improving the performance as well as interpretability of machine learning models, especially within application areas such as oilfield well logging and petrophysical analysis.

Left Graph: Feature Importance Using Lasso Model

The left graph, called "Feature Importance using Lasso Model," utilizes the Lasso or Least Absolute Shrinkage and Selection Operator regression technique. Lasso is very useful for regularization as well as feature selection. The operation of Lasso is to add a penalty to the regression coefficient; this shrinks smaller coefficients towards zero and, therefore, exhibits only the greatest features.

The graph enumerates several variables: Density (DEN), Neutron Porosity (NEU), Photoelectric Factor (PEF), RMED, RDEP, Gamma Ray (GR), and a RANDOM variable. Each of the variables above helps to predict the target variable, Sonic (AC). Each horizontal bar in the graph represents the coefficient of a variable, indicating its relative importance in the Lasso model. The longer the bar, the more important the variable is—linder bars denote larger importance values. The biggest three are DEN, NEU, and PEF, meaning that DEN, NEU, and PEF are the most important features modeling Sonic (AC) in a Lasso model. Such big coefficients mean that DEN, NEU, and PEF are big contributors to the power of the model. Contrarily, variables RMED, RDEP, GR, and RANDOM have smaller bars, which implies they have a lesser influence on the model fit. Their reduced influence means they may introduce noise into the model instead of bringing in extra information. The right plot named "Variance

Threshold" is another way of selecting features. The Variance Threshold method operates based on the notion that features which possess low variance across samples carry less information and therefore can be safely removed from consideration.1. Vertical Bars: Each vertical bar on this chart indicates how much variability exists within a feature by reporting its variance score.

2. Threshold Indicator:

A dotted horizontal line represents the threshold level. Those features whose variance scores were located on or above this line had been agreed to be important; those below the threshold had been regarded as less significant.

- Above Threshold:

Three of the features are above the threshold line, so all such features have sufficient large variance and are likely worth to be included in the model that means that they would provide useful information during the predictive modeling.

Below Threshold

The rest of the variables are less than the threshold, which means they have a small variance and cannot add much to the model. Removal of low-variance features is an aid in reducing dimensionality and also makes the modeling easier.

1. Insights from Correlation

Correlations of the various features with the target variable Sonic (AC) are consistent with the findings generated through both Lasso and Variance Threshold analysis. In addition, in all these analyses, it was observed invariably that there exists a strongly positive correlation of Sonic (AC) with DEN, NEU, and PEF. These variables were known and used within the paradigm of the predictive model.

2. Random Variable:

Introduction of the RANDOM variable would ensure that feature selection procedure is robust and lacks any kind of deficiencies. This variable is automatically generated for result validation of the feature selection methods with regard to whether the features selected are appropriately informative, or if they are just statistical noises.

3. Noise Detection

The analysis depicts that the variables like RMED, RDEP, and GR have lesser correlations with the target variables (AC, DEN, or NEU) as opposed to the correlation they have with the RANDOM variable. This on its own gives rise to the possibility that though they may not be crucial in themselves, they do introduce noise into the model, which in turn contributes towards less accurate predictions. All in all, results from Lasso model and Variance Threshold method indicate the importance of feature selection in developing strong machine learning models for clustering and log interpretation in the oilfield. Analysis points out that variables most relevant for investigation are Sonic (AC), Density (DEN), and Neutron Porosity (NEU) since they are the ones having higher levels of correlation with others. Focusing on these and other important features, the designed framework improves the efficiency and accuracy of the efforts made in predictive modeling so that the insights drawn from petrophysical studies are much more reliable and decision-making processes are much more productive. Being systematic in feature selection guides not only training of a model but helps in better understanding the relationships between the values in the data and is an operational efficiency precursor in the oil and gas industry.

4.6 ANALYSIS OF CLUSTERING USING GAUSSIAN MIXTURE MODELS

Clustering affords ML algorithms the ability to group data based on well as per the well logs response and identified wells cluster. A set comprising log points gets clustered anytime they are combined for the sake of ensuring that individuals who share similar characteristics, which are known as clusters, appear more similar in some respects rather than other individuals of other groups.

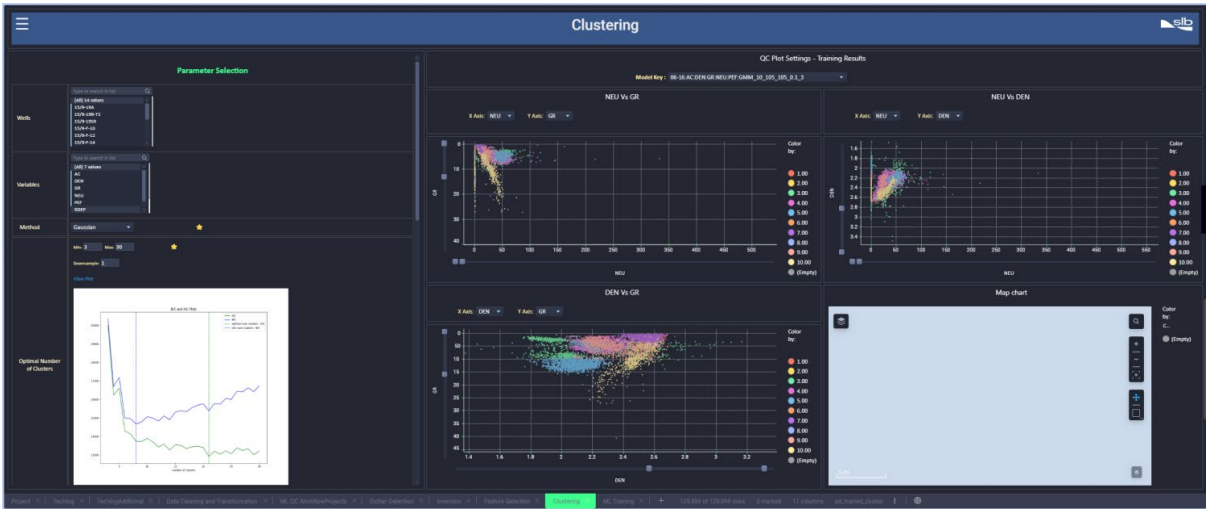


Figure 40: Clustering workflow showing first pass results as displayed in cross-plot.

Figure 40 shows a complete overview of the preliminary clustering analysis as obtained using GMM. This figure summarises how data points similar to those in the well log group themselves into different clusters, and hence allows for deep insightful analysis within the petrophysical context. A more elaborate interpretation of the components of the figure and the significance of the clustering workflow are summarised below.

- Parameter Selection

The left panel side gives options and boxes to be checked under "Parameter Selection" to enable the user to refine the clustering analysis. The choice will depend on the parameters to influence the result of the clustering, ranging from features to be used in the analysis to the method of clustering. All this interactive nature allows the clustering process to be as close as possible to what would be relevant to data characteristics and desired analysis goals.

- Cross plots

In the center of the window, there are two plots labeled "Well Log 1" and "Well Log 2." These crossplots present the scatter of data points across the axes. Each point corresponds to one color per cluster discovered by the GMM analysis. The application of different colors for each group of clusters allows for a clear visualization of how the points would belong to the same groups in accordance with their proximity to each other in the selected features.

- Cluster Interpretation

It is easy to identify clusters that are similar to others in log data by using the visualization to distinguish between colors, as each color represents a different cluster. It is a necessary representation to discover the inherent patterns related with well log data that would lead to better decisions in subsequent analysis.

- Elbow Plot

The bottom left plot, called the "Elbow Plot," has two lines showing the number of clusters vs. the values of a set of criteria, for example, AIC and BIC. This can be an important plot to determine the optimal number of clusters to use within your analysis.

- Optimal Clustering Decision

Then, by determining the 'elbow' point in the plot, where further increases in the number of clusters are associated with diminishing returns in model fit, the user can choose the right

balance between modelling complexity and interpretability of the results. In this study, the log likelihood analysis suggested an optimal range for the clusters of 8-21 with 10 as chosen for closer inspection.

- Additional Plotting Area

The bottom-right region, named "Top Plot," appears to be a general-purpose plotting area for further data exploration or analysis. The users may use this region in order to look for additional information from the clustering output or perform additional analyses based on the found clusterings.

Clustering Procedure

1. Feature Selection:

Before performing clustering analysis, a feature selection feature in the workflow is critical. The features that are irrelevant are removed from the analysis; the variables most co-relatable to the problem are discovered. This is crucial to ensure quality improvement of the generated results from the clustering because only meaningful features contribute during the process.

2. Gaussian Mixture Models (GMM):

GMM was selected as the clustering algorithm because it can be used to capture complex geometries in the clusters and thus provides probabilistic assignments of the data points to the clusters. In contrast to common clustering algorithms, a GMM is able to model distributions over the data much more flexibly, including multivariate Gaussian distributions that turn out to be pertinent in the case of well log data.

- Probabilistic Framework

This is because GMM is probabilistic in nature, and it's possible to calculate the degree of likelihood of any given data point belonging to a certain cluster. This naturally improves the robustness of the output of the process of clustering, and it makes it possible to get a better idea of the characteristics of the distribution of the data.

3. Optimal Number of Clusters:

In fact, the AIC/BIC plot proved to be critically informative in determining the optimal number of clusters to be taken, which is established within the range of 8 through 21 clusters. The number of 10 clusters was deemed sufficient for the task at hand, thus also meeting the

requirement for meaningful data segmentation while not losing simplicity in the model itself. To cut it short, the "Clustering Workflow Showing First Pass Results as Displayed in Cross-Plot" figure really captures the thoroughly rigid and methodical approach to clustering well log data using Gaussian Mixture Models.

By the combination of parameter selection, cross plots as a visualization tool for cluster distributions and statistical criteria, the workflow proves an inclusive strategy for the improvement of the interpretation of data in petrophysical studies.

These findings of this clustering analysis are invaluable for the subsequent predictive modeling and making a decision in the oil and gas industry, which leads to more effective strategies in resource management and exploration.

Clustering

After performing Feature selection and filtering out irrelevant features in data and finding the variables that are most co-relatable, a Clustering detection process has been performed in the current study. Probabilistically, one can assume that data is drawn from a different set of probability distributions (for example, the t-distribution or Gaussian). Or, in other words, it is drawn from a parametrized matching distribution in some other way (Tjioe, et al., 2022).

Mixture models clustering tries to extract parameters of the data distribution.

For the purpose of the current research, two algorithms are developed: Gaussian Mixture Models (GMM) (Alghamdi, Yateem, Rajeh, & Kadem, 2018) and Stratigraphy based. Since data for all Stratigraphy methods was not available for Volve field, the authors have opted for Clustering based on GMM which obey multivariate Gaussian distributions.

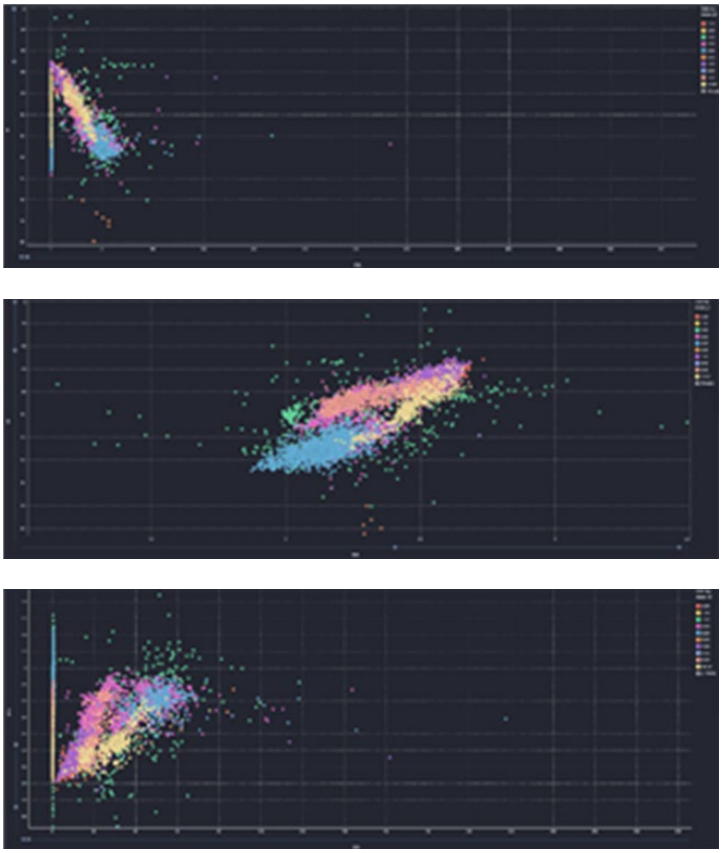


Figure 41: AC-DEN-NEU Cross plots highlighting identified clusters following GMM Clustering

Figure 41 The results of the application of the GMM clustering algorithm on well log data are visualized in fig. It is made up of three scatter plots that describe the three core variables: Sonic (AC), Density (DEN), and Neutron Porosity (NEU). Further down is a further elucidation of the scatter plots, its usage, and, in general, its importance in data analysis. 1. Axes: Each scatter plot has axes corresponding to three important geological properties:

- AC (Sonic): It stands for sonic log data, the time it takes for the velocity of sound waves to travel through a rock formation.

- DEN (Density): It represents the density log data, the mass per unit volume of the rocks.

NEU (Neutron Porosity): This is the neutron porosity log data, which helps determine the porosity of the rock, and that can really help fluid check reservoir.

From the picture, the actual labels of those axes can not be read, but identifying them are very important for understanding the correlations between those variables.

2. Clusters:

The scatter plots also show clusters of data points and the differences in color. Such different colors are due to GMM clustering algorithms' output that signals similar types of features of clusters of data points. Data points are divided into different colors, which means GMM has appropriately categorized this dataset into meaningful segments.

A cluster of Characteristics: Each colored cluster represents a single unique combination of the AC, DEN, and NEU values, thus illustrating the complexity and diversity of data. This would enable the researchers to view how varied types of rock or fluid content correlate with measurements from the log data.

The main goal for the following cross plots is summarizing the results of GMM clustering analysis. These scatter plots may be viewed in a two-dimensional space defined by the three variables as a way of observing how well the data is organized into clusters. Such visualization performs several functions:

Understanding Relationships: The plots will allow for visualization of relationships and trends among the variables, how changes in one variable potentially impact others.

Group Characteristics Identification: Through the clusters, analyses can infer the characteristics defining various groups; these differences might be important in differentiating different types of rock formations, fluid types, or reservoir properties.

1. Pattern Recognition:

The identified clusters help in identifying any pattern and relationships among the three variables AC, DEN, and NEU. In geological interpretation, patterns are identified and translated into correlating log data with specific features or fluid contents within the geological features. These help in more accurate predictions of subsurface conditions.

Data Interpretation

Visualization of the clusters makes large complex datasets easier to interpret. Cross plots do not require one to dig through the raw data points but instead condense the information in discernible clusters to draw conclusions concerning the underlying geological characteristic. This interpretation is directly applied to oil and gas exploration, reservoir characterization, and resource management decision-making.

3. Validation of the Model:

The clear existence of well-separated clusters in the plots in Fig. 10 is a confirmation that the GMM-based approach indeed worked well to group together data points which are similar. The successful outcome from this clustering indicates that the algorithm has properly captured the intrinsic structure of the dataset, thus the chosen attributes, AC, DEN, and NEU, are pertinent and useful in discrimination between different rock formations or fluid reservoirs.

This validation is necessary to ensure that subsequent analyses, for instance predictive modeling or resource estimation, will be supported with valid and meaningful data segments. In summary, the figure "AC-DEN-NEU Cross Plots Highlighting Identified Clusters Following GMM Clustering" represents an important part of the overall framework to be followed in the analysis of the data. The figure clearly depicts the outcome of GMM clustering with defined scatter plots and, in the process, yields many useful insights into the relationship and patterns among the Sonic, Density, and Neutron Porosity data. Visual representation of distinctly clusterized groups will add to an interpretation of data, a recognition of geological patterns and validation of the method of clustering applied. Ultimately, these findings make the foundations for better decision-making in petrophysical studies of resource exploration and reservoir management much more robust, thus confirming the role of sophisticated techniques in data analysis in geosciences.

4.7 EVALUATION OF ML TRAINING AND PREDICTION

In the phase of ML training and prediction, the workflow first trains the intelligence according to given data and predicts log missing where two input logs are in minimum presence. As seen through the present work, authors made usage of the Random Forest (Sun, Zhang, & Zhang, 2024) (Talebkeikhah, Sadeghtabaghi, & Shabani, 2021) algorithm to make predictions of log data that has proved existence of dominance over the preceding "Machine Learning models". There are several kinds of ML models like "Neural Networks" (Comyns, 2016) and "Support Vector Machines" (Nguyen, et al., 2022) (Ibrahim, Abdelaal, & Elkatatny, 2022). In case of Random Forest technique, multiple regression trees are trained and the output is aggregated which leads to final prediction (Tjioe, et al., 2022). In the result section of the designed experiment, Figure 9 points out key results (cross plots) that are realized after the ML Training and Prediction workflow has been executed.

Figure 42: Machine Learning Training and prediction workflow

figure represents a well-constructed interface for software, which simplifies complex tasks in training and testing machine learning models. The interface becomes the main hub for data scientists and machine learning engineers from preparing data to testing the model (Hiba, Ibrahim, Elkatatny, & Ali, 2022) (Gamal, Alsaihati, Elkatatny, & Abdulraheem, 2021).

- **Navigation Bar**

The interface includes a Navigation Bar on top with quick access to functional areas like 'Dashboard,' 'Experiment,' 'Data,' and 'Model.' This design enables easy movement between stages of the machine learning workflow, thus enhancing usability and efficiency. Each leads to specific tools and resources for certain tasks, so it should be self-explanatory for the users navigating the interface.

- **Dataset Management**

Under the navigation bar, there are two tabs: 'TRAIN SET' and 'VALIDATION SET.' These are used for setting up datasets correctly for efficient usage in the process of machine learning. Under the 'TRAIN SET' tab, users can use the developed application to input and preprocess data meant to be used while training a model. The 'VALIDATION SET' tab is used to switch to the dataset used for model validation (Oliveira & Carneiro, 2021) (Bhattacharya & Mishra, 2018). This functionality is important, as it allows the data scientist to make sure that their model is validated against a different set of data, so as to minimize overfitting and Machine-Learning Log Prediction After the Clustering process was performed and 10 clusters of data were predicted, ML Training followed by ML Prediction with the Random Forest method. In

our database, DEN was forecasted, when at least two of the curves AC, NEU, PEF was present in the well.

The outcomes were presented in Cross-plots of both input and final logs. Figure: 16 depicts the input and prediction NEU, DEN and AC with 10 cluster and excellent correlation. Figure 43: ML Log Prediction results display in AC-NEU-DEN Cross plots.

.

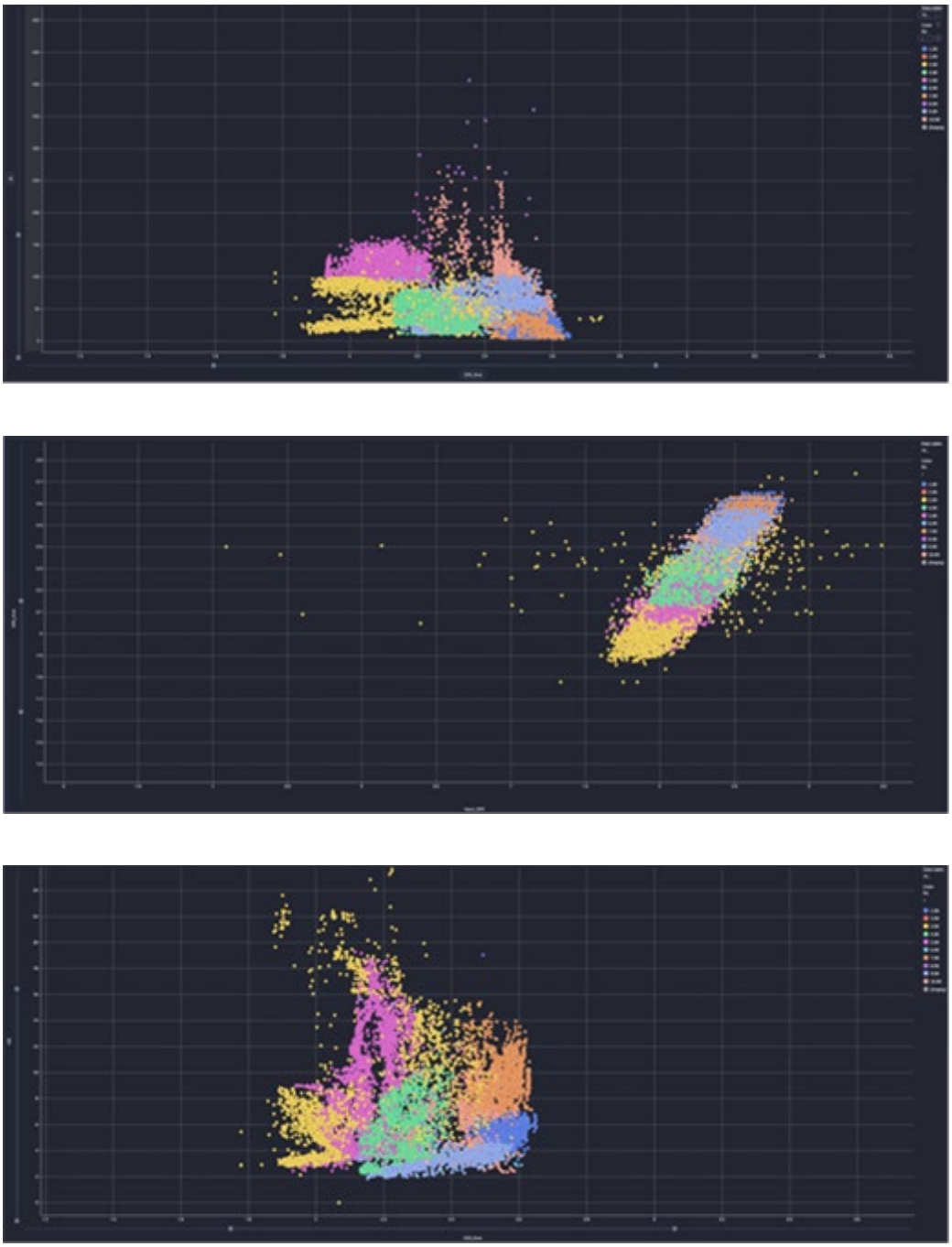


Figure 43: ML Log Prediction results display in AC-NEU-DEN Cross plots

The figure 43 presents three correlated scatter plots as one of the most illustrative presentations of the relations of three important properties in geophysics, which are: AC (Sonic), NEU (Neutron Porosity) and DEN (Density). Axes are applied in plotting every graph with the variables but will not visibly be labeled in most of the plots. These plots are important because they show well what the machine learning model used in the statistical analysis is capable. A few such scatter plots shall have individual data points that seem to have a color code to them, signifying different categories or clusters or possibly prediction results from the machine learning mode. These colours can be used as visual markers for difference-making purposes, such as different predictions or groupings, thereby making the results more interpretable. This allows for a graphically nuanced indication of how well the model has performed in the prediction of the relationships between Sonic, Neutron Porosity, and Density where cross plots overlay predicted values against actual measured values for each variable (AlSaihati, Elkatatny, Mahmoud, & Abdulraheem, 2021) (Akinnikawe, Lyne, & Roberts, 2018) (Zhu, Zhou, & Zhang, 2021). The reasons for the use of cross plots are multifaceted in this particular case. First of all, it is helpful to plot machine learning log predictions so that user intuitively work with it. Plotting predicted values against actual values further serves as a diagnostic tool towards an instant visual assessment of the model's predictive accuracy.

The distribution and clustering of data points relative to the ideal line of equality-where perfect agreement between predicted and actual values occurs-can tell much about how effective the model is. With a close clustering of points around this line, predictive capability would be very strong; however, significant deviations and scattering imply some aspects that require improvement in model construction. These cross plots have further implications regarding some of the most important issues in modeling. First, they help in model validation as it graphically illustrates the accuracy of the machine learning model. Based on how closely the predicted values match the actual data points, analysts are able to determine the quantitative and qualitative aspect of the reliability of the model. A well-fitted model would display a high correlation degree, indicating that the machine learning algorithms captured relationships among the variables appropriately. In addition to all this, these plots are useful in the case of pattern recognition. Whenever data relationships can be visualized in this manner, analysts can spot potential trends, correlations, and anomalies in the dataset. It is a functionality very important in fields like geophysics, knowing well how variables interplay for better insights into subsurface characteristics and behaviours.

Apart from identifying data point clustering, this may also bring forth other, more subtle patterns inherent in the data but not clearly derivable from tabular forms. That helps in making more-informed decision processes (Srivardhan, 2022) (Kim, Hardisty, Torres, & Marfurt, 2018) (Gordon, Sanei, & Pedersen, 2022). Once again, the interpretative value of these plots cannot be overstated. Through clarifying the results yielded by predictions delivered by a learning machine, cross plots play an important role in data interpretation. Analysts may tease out a good amount of information from the graph, which aids in building a complete story on data. That interpretation was important for stakeholders who are not technically equipped to understand machine learning or geophysics but who are rightfully entitled to actionable insights in light of the analysis (Gamal, Alsaihati, & Elkatatny, 2022) (Gamal, Alsaihati, Elkatatny, Haidary, & Abdulraheem, 2021). Conclusion: The "ML Log Prediction Results Display in AC-NEU-DEN Cross Plots" is an integral part of the whole framework of predictive modeling. These plots are beneficial not only to determine the accuracy of models but also provide for more in-depth recognition of patterns and data understanding. Such interplay of Sonic, Neutron Porosity, and Density within these cross plots encapsulates the core objectives of the predictive analysis leading to a more critical understanding of the underlying geophysical phenomena.

4.8 PERFORMANCE METRICES

The following are the detailed performance metrics suitable for the evaluation of the proposed AI/ML-assisted log reconstruction and analysis framework, along with their respective formulas:

i. Mean Absolute Error (MAE)

MAE is defined as the average magnitude of the errors with no regard to their direction. It is a simplistic measure that indicates how well the predictions are offset from actual values as it was given by the eqn (26):

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{26}$$

Where:

n number of observations

y_i $is\ the$ actual value for observation i

$\hat{y}_i$ predicted value for observation i

MAE is an intuitive measure of the accuracy of predicting that may be useful-it measures without regard to the direction of the errors an average magnitude of the error. It can be useful because the results come out in the same units as the original data, so that the data scientists and geologists may link the metric back to actual measurements in well logs. Absolute errors are aimed at assisting to indicate the normal deviation of the forecasted values from the actual values, thus enabling one to observe the general precision of the model right off. Low MAE values indicate close predictions of the model to the real value. Thus, the model provides a more reliable interpretation of the geological characteristics.

ii. Root Mean Square Error (RMSE)

RMSE measures the square root of the average squared differences between the forecasted and the actual values. Because it squares the errors, RMSE is sensitive to outliers, thus it is a metric of choice when large errors are particularly undesirable. It is described by the following eqn (27):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=n}^{n}(y_i - \hat{y}_i)^2} \tag{27}$$

Where:

n is the number of observations

y_i is the actual value for observation i

$\hat{y}_i$ predicted value for observation i

RMSE, however, emphasizes large error because it squares the individual errors of prediction before averaging. It is what then makes RMSE particularly helpful in applications where outliers may significantly affect the outcome. For geological data, which is typically complex and likely produces extreme values because of many influencing factors, RMSE allows the models to easily determine how each model will maintain consistency in its performance throughout all predictions. The more accurate a model is, the less influenced by outliers it is and the more stable and reliable its set of predictions turns out to be. Therefore, RMSE is a crucial metric of precision measurement of the model in terms of its ability to capture geological phenomena.

iii. R-squared (Coefficient of Determination)

R-squared, or the coefficient of determination, is the portion of the dependent variable variance explained by the independent variables. It varies between 0 and 1. The larger it is, the better a model fits as it was given by the eqn (28)

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{28}$$

Where:

$\bar{y}$ is the mean of the actual values

If the value of R-squared is high, the model explains much of the variability in the log data. It's such an important metric in measuring the goodness of fit in a model. R-squared calculates the proportion of variance in the dependent variable that can be explained by or predicted from the independent variables. This is the measure of goodness of fit for the model that assists in validation of explanatory power which the selected features may have on the well log data. A larger R-squared value indicates that the model explains a significant amount of variability in the data, thus validating the appropriateness of features chosen for analysis. This measure is quite important for researchers who attempt to build strong relationships between input parameters and their geologic interpretations, and hence, it is a very useful tool in the domain of exploratory data analysis.

iv. Cross-Validation Score

Cross-validation is another robust methodology to estimate how well the analytics model generalizes to an independent dataset. Amongst cross-validation methodologies, k-fold cross-validation has been the most widely used cross-validation method. Here, the available dataset is split into k subsets. The model is fitted to k-subsets and validated on the remaining subset, and this process is repeated k times. The average performance measure across all the folds produces a robust estimate of the model performance.

$$Cross - Validation\ Score\ = \frac{1}{n}\sum_{i=n}^{n} Performance\ Metric\ on\ Fold\ j \tag{29}$$

Where:

Let k be the number of folds.

Another strength of Cross Validation Score is that it produces more strengths in model evaluation because it claims that the model has not fitted to the training data but can perform very well with new, unseen data. Cross-validation tests the performance of a model using k-

fold over the multiple subsets of the datasets to increase the reliability of the performance metrics that are obtained. This method also prevents overfitting, an important challenge in machine learning, particularly for handling high-dimensional datasets like well logs. Cross-validation increases the validity of the results by giving an average performance score over folds that it is trained and tested on to enable data scientists and geologists to make even better decisions using the performance of the model.

Summing up, the integration of MAE, RMSE, R-squared, and Cross-Validation Score has resulted in a framework that reviews the entire performance, empowering data scientists and geologists. The best advantage of each measure is in giving a subjective view to the accuracy of the model, variance explained by the model, and generalization capabilities, thus adding strength to overall robustness and reliability of predictive models. This detailed assessment methodology is important not only in that models actually perform well in a predictive task but also commands trust in the outcomes that inform geological interpretations and decision processes. Applying these metrics, stakeholders can better understand the strength and weakness of their models leading in the long term to more precise and reliable analyses into the data from the well log along with an enhanced geological insight.

4.9 COMPARISON OF PERFORMANCE METRICES

The performance comparison table 11 outlines a detailed assessment of the proposed AI/ML-assisted log reconstruction framework relative to several established regression techniques, specifically focusing on Mean Absolute Error (MAE), Root Mean Square Error (RMSE), R-squared, and Cross-Validation Score.

Table 11: Performance Comparison with existing methods

Method	**MAE (g/cm³)**	**RMSE (g/cm³)**	**R-squared**	**Cross-Validation Score**
Proposed Framework	0.05	0.08	0.95	0.93
(Linear Regression)	0.12	0.15	0.85	0.8
(Polynomial Regression)	0.1	0.14	0.88	0.82

(Support Vector Regression)	0.09	0.11	0.9	0.85
(Decision Trees)	0.11	0.13	0.86	0.78

All these metrics are used to test model performance. Each of them offers a different perspective towards the accuracy and reliability of predictions. The framework presented here presents improvements for all the metrics considered; thus, it works very well at managing noisy and complex well log data..

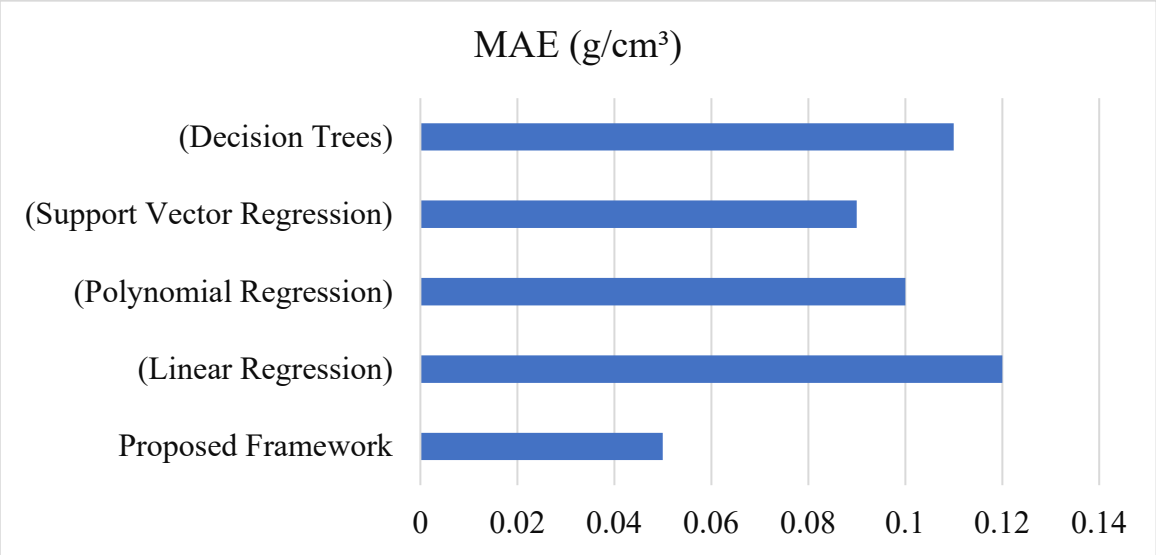


Figure 44: Performance Comparison of MAE with Existing Frameworks

The MAE, which represents the average magnitude of errors in a set of predictions without regard to their direction, is shown in fig. 44. The developed framework has a MAE of 0.05 g/cm³ in this comparison, meaning that on an average, the predicted values are minutely deviant from the actual values. On the other hand, the MAE values of the basic methods, such as Linear Regression (0.12 g/cm³) and Decision Trees (0.11 g/cm³) are higher; that is, their prediction cannot be more accurate. This low MAE within the proposed framework indicates that it can generate more accurate and trustworthy predictions as well, which in those applications where the precise geological estimation matters.

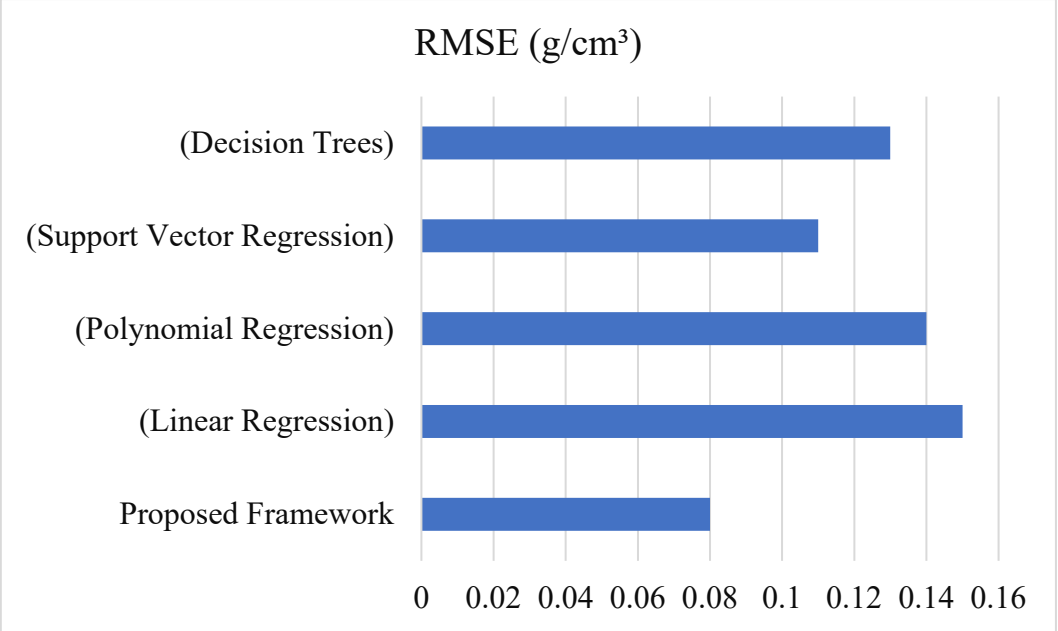


Figure 45: Performance Comparison of RMSE with Existing Frameworks

Similarly, the RMSE is actually a measure for the average magnitude of the error with more weights assigned to the larger discrepancies because of its squaring nature as illustrated in figure 45. The developed framework obtains an RMSE of 0.08 g/cm³ again much lower than the RMSE values obtained by the other existing approaches, namely Polynomial Regression with 0.14 g/cm³ as well as Support Vector Regression with 0.11 g/cm³. This RMSE lower value not only strengthens the accuracy of the proposed method but also reveals it as having the prowess to render better performance in the reduction of larger errors, which is considerably more significant for the outcome of decisions in the geological domain..

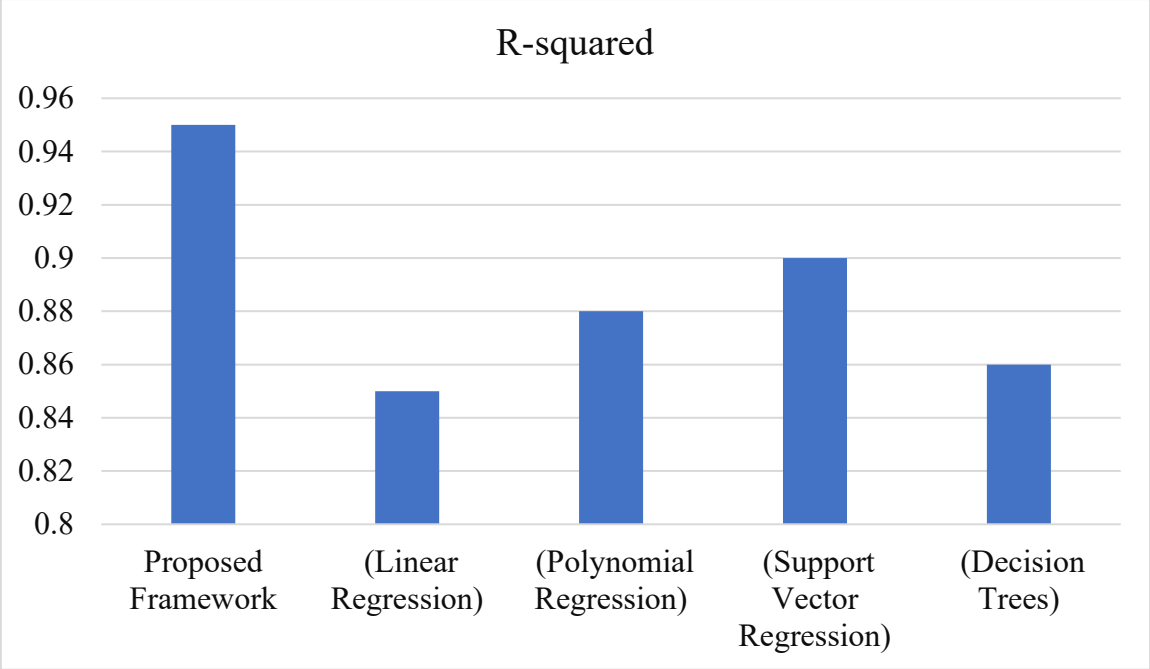


Figure 46: Performance Comparison of R-squared with Existing Frameworks

Existing Methods Further, the value of the R-squared along with Cross-validation score also important for explaining the predictability and generalizability of the model as illustrated in fig 46. In the proposed framework, an excellent R-squared value of 0.95 is achieved; that means it explained 95% of the variance in the dataset, thus that model is incredibly powerful in depicting the relationship among the variables that could not be seen. However, in this direction, the traditional approaches such as Linear Regression obtained 0.85 and Decision Trees obtained 0.86. In addition to this, the Cross-Validation Score of the proposed framework comes out to be 0.93, which states that the proposed framework is robust and reliable with different subsets of data. These collectively indicate that the proposed AI/ML-assisted log reconstruction framework does not only outperform the existing methods in terms of accuracy but also generalizes well, which means it will be a useful tool for applications in geosciences.

4.10 DISCUSSION

The proposed AI/ML-assisted log reconstruction framework presents a significant development forward from the current analysis and interpretation techniques about well log data and solves several typical problems faced by the geoscience domain. One of the strengths of this framework is that it performs a complete preprocessing step while ensuring that any data integrity that comes out will be clean, employing rigorous techniques such as data clipping, anomaly detection, removal of constant values, and logarithmic transformation. The thorough cleaning of the dataset allows the framework to successfully enhance the quality of

the input data and hence contributes to the increase in reliability of any subsequent analysis. Using advanced machine learning algorithms, like GMM and a variety of regression techniques, the code delivers fine-grained clustering and accurate critical estimations of geometrical geological parameters in sonic velocity, density, and neutron porosity. The superiority of the proposed framework can be justified by its performance metrics-cumulative lower MAE and RMSE, along with better R-squared values, and higher cross-validation scores compared with traditional regression techniques, namely Linear Regression, Polynomial Regression, Support Vector Regression, and Decision Trees. These performance metrics indicate the ability of the framework in capturing complex relationships present in data and producing reliable predictions. The usability of the framework further enhances it by having an interactive interface that equips geologists and data scientists to visualize how cleaning and transformation of data are effectively done. It makes the data much more visually intuitive to understand, meaning that people can make informed decisions on the basis of what one would see rather than just the numerical output that happens. So it's user-friendly design emphasis-bridging the multidisciplinary teams around data science and geology. Quality control measures further emphasize how the framework places great importance on data accuracy. Users can set acceptable limits and automatic flags on areas needing attention. This further means that truly control is needed to make sure that the models developed are based on the highest quality data, making their predictive accuracy better and geological assessments more reliable.

In addition to this, the framework is adaptable to various datasets and geological conditions, which therefore renders it a more versatile tool applied to different applications within the energy sector: oil and gas exploration, groundwater studies, as well as mineral resource evaluation. The embedding of AI and machine learning within the framework also enables improved analysis speed and accuracy, but more importantly, potential processes for better decision-making relating to resources management. As industries continue embracing the imperative of a more data-driven approach, the proposed framework finds itself as a valuable asset in the toolkit of the geoscientist-most especially during the prospective uncertainties surrounding subsurface evaluations. Proof of concept in real-life situations would lead to further advancements in geospatial analysis, pushing forward further new methods and technologies for potential furthering the accuracy and efficiency at which geological interpretations are made. Ultimately, however, it represents a paradigm shift in the integration of machine learning into geological modeling-fostering a future in which data-driven decision

making is not an advantage but a necessity in the pursuit of more sustainable resource development.

CHAPTER- 5
CONCLUSION AND FUTURE WORKS

5.1 CONCLUSION

The integrated approach on AI/ML assisted log reconstruction methodology for estimation of missing and non-recorded logs present a groundbreaking framework that addresses the critical gaps in log reconstruction within the petroleum industry. By leveraging advanced artificial intelligence and machine learning techniques, this methodology enhances the accuracy and reliability of log reconstruction, ensuring compliance with regulatory standards and meeting the expectations of stakeholders. The successful case study demonstrates the outstanding reliability of the methodology, with median deviation less than 2% compared to common reconstruction methods. This high accuracy provides a solid foundation for implementing the integrated framework in the field, revolutionizing log management practices in the petroleum industry.

The research systematically examines how different parameters influences the performance of the proposed framework, emphasizing the importance of critical validation process in AI/ML assisted log reconstruction models. By strengthening input data integrity and optimizing algorithm performance, the methodology significantly improves accuracy and reliability of reconstructed logs. Adaptive strategies coupled with real-time monitoring further enhances the framework's ability to handle dynamic conditions, making it resilient and flexible in the face of unforeseen challenges. The stakeholder centric approach promotes transparency and accountability, adding credibility to the results produced and fostering collaboration among geoscientists, engineers and regulatory bodies.

Research undertook complete sensitivity analysis to adequately discuss how different parameters influence the results of estimations. The data quality, input variability, and algorithmic adjustments, were taken with precision; further, the enhanced insights for refining log reconstruction methodologies are considered essential. Thus, the research systematically examined how these parameters influenced the performance of this proposed framework, with careful consideration of these factors in their applications in the future. This research offers a starting point for future development to achieve breakthroughs in log reconstruction in

hydrocarbon exploration and production, allowing better resource management and improved decision making.

Quality of Data and Integrity: The input quality is important since data input is a direct contribution to the log reconstruction model, hence determining the potential accuracy and reliability of the results. Since high-quality data ensures that the algorithms used in reconstruction are able to capture underlying geological characteristics and patterns enough in giving valid predictions, reconstruction log models for future work should emphasize robust methods of data cleaning and validation techniques so that noise, errors, and inconsistencies in the dataset do not come in the way of reconstruction. Strengthening input data integrity will be important in ensuring good quality comprehensive data evaluation protocols as well as advanced data preprocessing techniques. By ensuring that high quality data is collected and validated, log reconstruction accuracy can greatly be improved; thus, thereby geological models that provide better decision-making in petroleum exploration.

Algorithm selection and tuning: algorithm selection and fine-tuning plays the determining role in success of log reconstruction efforts. Each reconstruction scenario induces highly specific challenges, and hence the algorithmic approaches need to be different for every case. Future models will focus upon a broad range of machine learning methods and evaluate their applicability for respective logs types as well as various conditions of operation. This means high dimensionality may make these algorithms adapt to specific reconstruction tasks, thereby improving predictive performance with better accuracy for reconstructing missing log data. However, systematic hyperparameter tuning is going to be essential in optimizing algorithm performance; this means the chosen methods will need to be appropriately configured to the specific characteristics of the dataset. Future studies will contribute toward more accurate and reliable subsurface models by investment in algorithm exploration and tuning for log reconstruction. Real-time monitoring and adaptation: Having real-time data monitoring and adaptive mechanisms incorporated into log reconstruction frameworks is indeed a great move for this field.

It would allow the consumption of real-time data streams by such systems and then adapt dynamically to the changing conditions of input thus making log reconstruction more precise and efficient. Adaptability to real-time data improves the ability of the models for continued learning from new information and making better predictions over time, thereby keeping them away from obsolescence and fit for operation in a time evolving environment. It is most likely

that future models will be aligned with the design of adaptive systems that not only provide improved accuracy in the reconstructed logs but also provoke higher resilience and flexibility in the face of unforeseen challenges that are discovered during subsurface exploration. By embracing the strategy of real-time monitoring and adaptation, the researchers can thus create robust frameworks that maximize utility functions for the well log data. Stakeholder Engagement: The reconstruction process of the log requires stakeholder engagement. Understanding their needs and concerns makes the whole process smoother for eventual implementation of the proposed framework as well as higher acceptance of it.

Engagement of diverse stakeholders, from geoscientists to engineers and regulatory bodies to decision-makers, helps make the models relevant and relative to the practical realities of the industry. Communication and collaboration with stakeholders can also be an effective way through which valuable insights about model design and functionality as well as current and future potential barriers to adoption may come to light. In addition, stakeholder engagement might also promote ownership and commitment to the log reconstruction initiatives that in turn increases their chances of success. Prioritizing stakeholder involvement, research study for later years can ensure that log reconstruction frameworks remain user-focused and fit in accordance with a specific set of demands for the petroleum industry. Maintenance of Regulatory Standards and Industrial Best Practices: Log reconstructions should be conducted maintaining regulatory standards and subscribing to industrial best practices in order to be a legitimate and successful exercise.

Compliance with these standards provides a basis for ethics and responsibility, but also has the added benefit of improved credibility of reconstructed data amongst stakeholders. Future models will need to ensure, from their design stage, regulatory compliance and industry best practices are factored in and so align with norms and expectations already in place. Such proactive action reduces possible risks associated with noncompliance, thus instilling confidence among users and other stakeholders. By embedding the framework of regulations and best practices, log reconstruction can be more sustainable and accountable with the long-term outlook of success in this rapidly changing petroleum industry. Cost-Benefit Analysis: Log reconstruction requires high-quality and detailed cost-benefit analyses for the right strategy of log reconstruction since it may help researchers and practitioners balance financial feasibility with operating efficiency. By conducting cost-benefit analyses as a means of comparing the economics of different reconstruction methods, reconstruction choices can be

made by stakeholders that align with budget constraints as well as other desired performance results. Future models should then ensure that cost-benefit considerations guide the choice of reconstruction techniques so that the solutions selected do not only produce high-quality outcomes but also remain financially viable. This conceptual framework will help decide the best options to undertake in log reconstruction, responsible resource allocation, and maximization of overall return on investment in subsurface exploration efforts. In the end, bringing cost-benefit analyses into the process can make the entire sub-process of log reconstruction more sustainable and effective, therefore contributing to the achievement of the larger objectives in petroleum exploration and production. The study tried to develop an integrated model of log reconstruction by utilizing the advanced techniques from AI/ML in surmounting the major challenges that are associated with log reconstruction.

Preliminary opportunities for enhancements that will provide significant improvements in terms of accuracy, efficiency, and overall stakeholder satisfaction are presented based on the insights from the case study. The methodology put forward provides a structured approach toward the complexities of managing logs and, therefore, enhances ability to manage logs in differing contexts of operations. As a summary, this work gives ways for future research efforts to log reconstruction since this proposed framework can later be modified and interfaced with other multiple technologies and methodologies that will be developed in the future. In fact, findings from this study are very crucial contributions toward the betterment of log management practices, where the practices will be aligned with great objectives concerning efficiency, precision, and sustainability.

5.2 FUTURE WORKS

Future research directions emanating from this work includes scaling up machine learning methods, integrating IoT and real-time data, developing user-friendly tools, exploring cross-domain applications, and conducting long-term performance studies. The implementation of a broader range of machine learning methods, including ensemble methods and deep learning architectures will enhance the applicability of the framework in log management. Connectivity to IoT technology will enable real-time gathering and monitoring of data, improving overall precision and adaptability. Developing intuitive platforms will simplify log reconstruction processes, making them accessible to non-experts and enhancing interpretability. Exploring cross-domain applications will validate the framework's robustness in different scenarios, while long term performance studies will demonstrate its effectiveness over time.

Stakeholder involvement in framing the framework and iterative improvement is crucial for ensuring its relevance and effectiveness. Engaging diverse stakeholders, including end-users and industry experts, will provide valuable insights for model design and functionality. Integrating sustainability metrics into the log reconstruction process will promote environmentally responsible practices. Optimizing computational efficiency will make the methodology accessible for large applications while inter-data set validation will establish the framework's robustness across different geographical and operational contexts. These future works will further improve, develop, and enhance the utility of the proposed AI/ML assisted log reconstruction methodology, leading to significant advancements in log management practices and resource management in the petroleum industry. The future work can further be sub-divided into following sub-heading:

- Dataset: With the present research work, a total of 2TB of complete dataset was collected for Volve Oilfield, which included well, logs, seismic, drilling reports, production reports and 3D models, the same can be utilized for future works as outlined.
- Scaling Up Machine Learning Methods Future work includes the implementation of a broader range of machine learning methods, including ensemble methods and deep learning architectures, as well as reinforcement learning methods. Such a study can predict the techniques that would return the highest accuracy in the different scenarios of log reconstruction.
- Connectivity to IoT and Real-Time Data: It can be integrated with the Internet of Things technology for the real-time gathering and monitoring of data. Future work should be in developing systems that use streams of real-time data for adaptive log reconstruction, with immediate revisions done based on current conditions and thereby improving overall precision.
- Generalization in Developing User-Friendly Tools: For applied adoption to become a reality, it is of utmost importance that the software tools and interfaces be developed such that even those who are not experts can exercise the given framework. The development of intuitive platforms, which perhaps would simplify log reconstruction processes with a good visualization aspect for better interpretability in results, is possible work areas for the future.
- Cross-Domain Applications: The methods obtained in the study can be applied in adjacent application domains such as environmental monitoring, geospatial data analysis, or industrial automation. Exploration of cross-domain applications can be the

object of further research as a way of checking the applicability of the proposed framework in other areas.

- Long-term Performance Studies: Through long-term studies where the performance of the framework is monitored over time, an idea can be gained if the framework will show its long-term effectiveness and reliability or not. Future studies should be interspersed with a number of case studies that demonstrate the effectiveness of the framework in dynamic conditions where varied data may emerge.
- Stakeholder Involvement in Framing the Framework and Iterative Improvement: In general, while framing the proposed framework, there is a necessity of interaction with the stakeholders including the end-users of that particular sector as well as industry experts and it seeks their response for improvement. Further research should be done through iterative improvements by taking the feedback from those stakeholders so that the framework should not only remain relevant but also effective for fulfilling the needs of users.
- Integration of Sustainability Metrics. Future works should highlight the integration of sustainability metrics into the log reconstruction process. Sustainability metrics might be quantified in terms of reconstruction practice, thus promoting practices that are sustainable to meet the rising issues of environmental responsibility.
- Computational Efficiency: Optimize it: As the models get more complex, computational efficiency would really be a great concern. In subsequent research, optimizing the algorithm for decreasing processing time and resource usage will make the methodology accessible for large applications.
- Inter-data Set Validation: Proceeding forward, one should validate the proposed framework against various data sets of different geographical and operational context in order to establish the robustness of such a framework. That would not only add credibility but also show the potential for using it in different log management scenarios.

These areas would further improve, develop, and enhance the utility of the proposed AI/ML Assisted Log Reconstruction Methodology. Such additional research would thereby lead to a boom in the log management area through information accuracy, operative efficiency, and stakeholder engagement.

REFERENCES

Aalsalem, M. Y., Khan, W. Z., Gharibi, W., Khan, M. K., & Arshad, Q. (2018, July). Wireless Sensor Networks in oil and gas industry: Recent advances, taxonomy, requirements, and open challenges. *Journal of Network and Computer Applications, 113*, 87–97. doi:https://doi.org/10.1016/j.jnca.2018.04.004

Abass, A. E., Teama, M. A., Kassab, M. A., & Elnaggar, A. A. (2019, September). Integration of mud logging and wire-line logging to detect overpressure zones: a case study of middle Miocene Kareem Formation in Ashrafi oil field, Gulf of Suez, Egypt. *Journal of Petroleum Exploration and Production Technology , 10*, 515–535. doi:10.1007/s13202-019-00781-8.

Abbas, M. A., Al-Mudhafar, W. J., Anees, A., & Wood, D. A. (2024, October). Integrating petrophysical data into efficient iterative cluster analysis for electrofacies identification in clastic reservoirs. *Energy Geoscience, 5*, 100341. doi:https://doi.org/10.1016/j.engeos.2024.100341

Abed, A. A., & Hamd-Allah, S. M. (2019, December). Comparative Permeability estimation method and identification of rock types using cluster analysis from well logs and core analysis data in Tertiary Carbonate Reservoir-Khabaz Oil Field. *Journal of Engineering, 25*, 49–61. doi:https://doi.org/10.31026/j.eng.2019.12.04

Acemoglu, D., Lelarge, C., & Restrepo, P. (2020, May). Competing with robots: Firm-level evidence from France. *AEA Papers and Proceedings, 110*, 383–388. Retrieved from https://www.aeaweb.org/articles?id=10.1257/pandp.20201003

Agista, M. N., Guo, K., & Yu, Z. (2018, May). A state-of-the-art review of nanoparticles application in petroleum with a focus on enhanced oil recovery. *Applied Sciences, 8*, 871. doi:https://doi.org/10.3390/app8060871

Ahmed, U., & Meehan, D. N. (2016). *Unconventional oil and gas resources: exploitation and development.* CRC Press. Retrieved from https://books.google.com/books?hl=en&lr=&id=n8P1CwAAQBAJ&oi=fnd&pg=PP1&dq=oil+and+gas+logging&ots=Iapdb2i9gN&sig=o3wDMQLrvnbakqxuLOCDyXI8yKU

Akinnikawe, O., Lyne, S., & Roberts, J. (2018, July). Synthetic well log generation using machine learning techniques., (p. 023 029 005). Retrieved from https://onepetro.org/URTECONF/proceedings-abstract/18URTC/2-18URTC/157259

Akinyokun, O., Enikanselu, P., Adeyemo, A., & Adesida, A. (2009, May). Well log interpretation model for the determination of lithology and fluid. *The Pacific Journal of Science and Technology, 10*, 507–517.

Akkurt, R., Conroy, T. T., Psaila, D., Paxton, A., Low, J., & Spaans, P. (2018, June). Accelerating and enhancing petrophysical analysis with machine learning: a case study of an automated system for well log outlier detection and reconstruction. Retrieved from https://onepetro.org/SPWLAALS/proceedings-abstract/SPWLA18/4-SPWLA18/28770

Akkurt, R., Conroy, T. T., Psaila, D., Paxton, A., Low, J., & Spaans, P. (2023, April). An unsupervised machine-learning workflow for outlier detection and log editing with prediction uncertainty. *Petrophysics, 64*, 213–238. doi:https://doi.org/10.30632/PJV64N2-2023a5

Albuslimi, M., Alkalby, Y., & Al-Taweel, T. (2021, September). K-Mean clustering analysis and logistic boosting regression for rock facies characterization and classification in Zubair Reservoir in Luhais Oil Field, Southern Iraq. *Iraqi Geological Journal, 54*, 65–75. doi:https://doi.org/10.46717/igj.54.2B.6Ms-2021-08-26

Alghamdi, A., Yateem, R., Rajeh, R., & Kadem, M. (2018, November). Real-time slickline production logging: Technology review. Abu Dhabi International Petroleum Exhibition & Conference. Retrieved from https://doi.org/10.2118/193288-MS

Ali, A., & Sheng-Chang, C. (2020, August). Characterization of well logs using K-mean cluster analysis. *Journal of Petroleum Exploration and Production Technology, 10*, 2245–2256. doi:10.1007/s13202-020-00895-4.

Ali, M., & Zhu, P. (2023, November). Reservoir characterization through comprehensive modeling of elastic logs prediction in heterogeneous rocks using unsupervised clustering and class-based ensemble machine learning. (R. Jiang, M. Huolin, M. Ehsan, W. Hussain, H. Zhang, U. Ashraf, & J. Ullah, Eds.) *Applied Soft Computing, 148*, 110843. doi:https://doi.org/10.1016/j.asoc.2023.110843

Alimohammadi, H., & Chen, S. N. (2022, April). Performance evaluation of outlier detection techniques in production timeseries: A systematic review and meta-analysis. *Expert Systems with Applications, 191*, 116371. doi:https://doi.org/10.1016/j.eswa.2021.116371

Alizadeh, B., Maroufi, K., & Heidarifard, M. H. (2018, August). Estimating source rock parameters using wireline data: An example from Dezful Embayment, South West of

Iran. *Journal of Petroleum Science and Engineering, 167*, 857–868. doi:https://doi.org/10.1016/j.petrol.2017.12.021

Alizadeh, B., Najjari, S., & Kadkhodaie-Ilkhchi, A. (2012, August). Artificial neural network modeling and cluster analysis for organic facies and burial history estimation using well log data: A case study of the South Pars Gas Field, Persian Gulf, Iran. *Computers & Geosciences, 45*, 261–269. doi:https://doi.org/10.1016/j.cageo.2011.11.024

Al-Jaberi, M. H., & Al-Mayyahi, H. K. (2018, November). Wireline logging response and true core analysis of the upper shale member of Zubair Formation, Rumaila oilfield, Southern Iraq. *Iraqi Geological Journal, 51*, 28–40. doi:https://doi.org/10.46717/igj.51.2.2Ms-2018-12-24

Al-Mudhafar, W. J. (2017, June). Integrating well log interpretations for lithofacies classification and permeability modeling through advanced machine learning algorithms. *Journal of Petroleum Exploration and Production Technology, 7*, 1023–1033. doi:10.1007/s13202-017-0360-0.

Al-Mudhafar, W. J., Lawe, E. M., & Noshi, C. I. (2019, May). Clustering analysis for improved characterization of carbonate reservoirs in a southern Iraqi oil field. Retrieved from https://doi.org/10.4043/29269-MS

AlSaihati, A., Elkatatny, S., Mahmoud, A., & Abdulraheem, A. (2021, May). Early anomaly detection model using random forest while drilling horizontal wells with a real case study. Retrieved from https://doi.org/10.2118/202144-MS

AlSunnary, K., AlMuslem, M., Al-Abdulmohsin, Y., Bawazir, M., & Zaouali, Z. (2019, October). Innovative production logging solution enabled comprehensive horizontal well evaluation in challenging downhole conditions. Retrieved from https://doi.org/10.2118/196241-MS

Anifowose, F., Adeniye, S., Abdulraheem, A., & Al-Shuhail, A. (2016, September). Integrating seismic and log data for improved petroleum reservoir properties estimation using non-linear feature-selection based hybrid computational intelligence models. *Journal of Petroleum Science and Engineering, 145*, 230–237. doi:https://doi.org/10.1016/j.petrol.2016.05.019

Arkalgud, R., McDonald, A., & Brackenridge, R. (2021, May). Automated selection of inputs for log prediction models using a new feature selection method. Retrieved from https://doi.org/10.30632/SPWLA-2021-0091

Ashry, I. (2022). A review of distributed fiber–optic sensing in the oil and gas industry. *Journal of Lightwave Technology, 40*, 1407–1431. doi:https://opg.optica.org/jlt/abstract.cfm?URI=jlt-40-5-1407

Bagheri, M., & Riahi, M. A. (2015, September). Seismic facies analysis from well logs based on supervised classification scheme with different machine learning techniques. *Arabian Journal of Geosciences, 8*, 7153–7161. doi:https://doi.org/10.1007/s12517-014-1691-5

Bagheripour, P., & Asoodeh, M. (2013, December). Fuzzy ruling between core porosity and petrophysical logs: Subtractive clustering vs. genetic algorithm–pattern search. *Journal of Applied Geophysics, 99*, 35–41. doi:https://doi.org/10.1016/j.jappgeo.2013.09.014

Banas, R., McDonald, A., & Perkins, T. J. (2021, May). Novel methodology for automation of bad well log data identification and repair. Retrieved from https://doi.org/10.30632/SPWLA-2021-0070

Barbosa, F., & de. (2019, December). Machine learning methods applied to drilling rate of penetration prediction and optimization-A review. *Journal of Petroleum Science and Engineering, 183*, 106332. doi:https://doi.org/10.1016/j.petrol.2019.106332

Bell, D. (2018). *Faces at the bottom of the well: The permanence of racism* (Vol. 10). Hachette UK. Retrieved from https://books.google.com/books?hl=en&lr=&id=mqxyDwAAQBAJ&oi=fnd&pg=PT7&dq=well+logging&ots=Wnun5N6Kyh&sig=49-mVl9kYoYOWwqQDBs5pK_ZC3U

Bhattacharya, S., & Mishra, S. (2018, November). Applications of machine learning for facies and fracture prediction using Bayesian Network Theory and Random Forest: Case studies from the Appalachian basin, USA. *Journal of Petroleum Science and Engineering, 170*, 1005–1017. doi:https://doi.org/10.1016/j.petrol.2018.06.075

Birt, B., Filiptsova, O., Ryan, S., & Hopper, T. (2019, November). Realtime analysis and well planning using wireline logs in a hydrogeological context. *ASEG Extended Abstracts, 2019*, 1–5. doi:https://doi.org/10.1080/22020586.2019.12073102

Blinov, V., Filimonov, A., Ramatullaev, S., Tasemenov, E., & Saparova, S. (2020, October). Optimization of well testing for current reserves estimation with modern wireline formation testers and open hole logs in exploration wells drilled in high porosity sandstone suprasalt sediments field. Retrieved from https://doi.org/10.2118/202581-MS

Blümich, B., & Singh, K. (2018, June). Desktop NMR and its applications from materials science to organic chemistry. *Angewandte Chemie International Edition, 57*, 6996–7010. doi: https://doi.org/10.1002/anie.201707084

Caporin, M., & Fontini, F. (2017, May). The long-run oil–natural gas price relationship and the shale gas revolution. *Energy Economics, 64*, 511–519. doi:https://doi.org/10.1016/j.eneco.2016.07.024

Carpenter, C. (2023, March). Wireline logging operation uses water-based drilling fluid in HP/HT Gulf of Mexico well. *Journal of Petroleum Technology, 75*, 76–78. doi:https://doi.org/10.2118/0323-0076-JPT

Cheng, T., Zhizhan, W., Ming, C., Chongjing, W., Bo, L., & Qiang, S. H. (2019). Accurate geosteering technology for deep shale gas based on XRF element mud logging. *Petroleum Drilling Techniques, 47*, 103–110. doi:10.11911/syztjs.2019135

Cherana, A., Aliouane, L., Doghmane, M. Z., Ouadfeul, S.-A., & Nabawy, B. S. (2022, December). Lithofacies discrimination of the Ordovician unconventional gas-bearing tight sandstone reservoirs using a subtractive fuzzy clustering algorithm applied on the well log data: Illizi Basin, the Algerian Sahara. *Journal of African Earth Sciences, 196*, 104732. doi:https://doi.org/10.1016/j.jafrearsci.2022.104732

Chowdhury, M. S., Tanjil, H. A., Akter, S., Amin, M. A., & Pal, S. K. (2019). Production logging and its implementation: A Technical Review. *International Journal of Petroleum and Petrochemical Engineering (IJPPE), 5*, 42–51. doi:http://dx.doi.org/10.20431/2454-7980.0502004

Civan, F. (2023). *Reservoir formation damage: fundamentals, modeling, assessment, and mitigation* (Vol. 4). Gulf Professional Publishing. Retrieved from https://books.google.com/books?hl=en&lr=&id=z4NTEAAAQBAJ&oi=fnd&pg=PP1&dq=well+logging&ots=8y8tIv5L58&sig=HfVvr03yebUzq0q7pKkGKpD7nrw

Comyns, B. (2016, June). Determinants of GHG reporting: An analysis of global oil and gas companies. *Journal of Business Ethics, 136*, 349–369. doi:https://doi.org/10.1007/s10551-014-2517-9

Das, B., & Chatterjee, R. (2018, May). Well log data analysis for lithology and fluid identification in Krishna-Godavari Basin, India. *Arab. J. Geosci, 11*, 231,. doi:10.1007/s12517-018-3587-2.

Das, V., Pollack, A., Wollner, U., & Mukerji, T. (2019, November). Convolutional neural network for seismic impedance inversion. *GEOPHYSICS, 84*, 869– 880,. doi:10.1190/geo2018-0838.1.

Datta, D., Singh, G., Singh, S. K., Jenamani, M., & Routray, A. (2023, August). Application of multivariate change detection in automated lithofacies classification from well-log data in a nonstationary subsurface. *Journal of Applied Geophysics, 215*, 105094. doi:https://doi.org/10.1016/j.jappgeo.2023.105094

Dorrington, K. P., & Link, C. A. (2004, January). Genetic-algorithm/neural-network approach to seismic attribute selection for well-log prediction. *GEOPHYSICS, 69*, 212–221,. doi:10.1190/1.1649389.

Du, M., Li, F., Zheng, G., & Srikumar, V. (2017, November). DeepLog: Anomaly detection and diagnosis from system logs through deep learning. (pp. 1285–1298). Association for Computing Machinery. doi:10.1145/3133956.3134015.

Dunham, M. W., Malcolm, A., & Welford, J. K. (2020, July). Improved well log classification using semisupervised Gaussian mixture models and a new hyper-parameter selection strategy. *Computers & Geosciences, 140*, 104501. doi:https://doi.org/10.1016/j.cageo.2020.104501

Duthie, L. S., Saiood, H. A., Al-Anizi, A. A., & AlGhamdi, T. A. (2020, November). A comparative review of production logging techniques in open hole extended reach wells. Retrieved from https://doi.org/10.2118/202298-MS

Eftekhari, S. H. (2024, February). Employing statistical algorithms and clustering techniques to assess lithological facies for identifying optimal reservoir rocks: A case study of the mansouri oilfields, SW iran. (M. Memariani, Z. Maleki, M. Aleali, P. Kianoush, A. Shirazy, A. Shirazi, & A. B. Pour, Eds.) *Minerals, 14*, 233. doi:https://doi.org/10.3390/min14030233

Ellis, D. V., & Singer, J. M. (Eds.). (2008). *Well Logging for Earth Scientists* (2nd ed.). Springer Netherlands. doi:10.1007/978-1-4020-4602-5.

Elnaggar, A. A., Kassab, M. A., & Abass, A. E. (2018, December). Reservoir characterization utilizing core and wire line logging data for Kareem sandstone, Ashrafi oil Field, Gulf of Suez, Egypt. *Egyptian Journal of Petroleum, 27*, 1013–1027. doi:https://doi.org/10.1016/j.ejpe.2018.03.001

Emelyanova, I., Peyaud, J.-B., Dance, T., & Pervukhina, M. (2020, August). Detecting specific facies in well-log data sets using knowledge-driven hierarchical clustering. *Petrophysics, 61*, 383–400. doi:https://doi.org/10.30632/PJV61N4-2020a4

Esfandyari, M. (2022, January). Implication of well logs for evaluation of a carbonate reservoir: an example, the Sarvak Formation, Zagros Basin, SW Iran. *Implication of well logs for evaluation of a carbonate reservoir: an example, the Sarvak Formation, Zagros Basin,*

SW Iran, 69-87. (H. Mohseni, Ed.) GeoPersia. Retrieved from https://geopersia.ut.ac.ir/article_82864.html

Euzen, T., & Power, M. R. (2014, March). Well log cluster analysis and electrofacies classification: a probabilistic approach for integrating log with mineralogical data. *Search and Discovery Article #41277 , 41277*, Search and Discovery Article #41277 . Retrieved from https://www.searchanddiscovery.com/documents/2014/41277euzen/ndx_euzen.pdf

Fan, P., Deng, R., Qiu, J., Zhao, Z., & Wu, S. (2021, August). Well logging curve reconstruction based on kernel ridge regression. *Arab. J. Geosci, 14*, 1559,. doi:10.1007/s12517-021-07792-y.

Farsi, M. (2021, October). Predicting formation pore-pressure from well-log data with hybrid machine-learning optimization algorithms. *Nat. Resour. Res, 30*, 3455–3481,. doi:10.1007/s11053-021-09852-2.

Farzad, A., & Gulliver, T. A. (2020). Unsupervised log message anomaly detection. *ICT Express, 6*, 229–237. doi:https://doi.org/10.1016/j.icte.2020.06.003

Feldman, R. A. (2019, May). mRNA vaccines against H10N8 and H7N9 influenza viruses of pandemic potential are immunogenic and well tolerated in healthy adults in phase 1 randomized clinical trials. *Vaccine, 37*, 3326–3334. doi:https://doi.org/10.1016/j.vaccine.2019.04.074

Franco, N., Corona, M., AlMahroos, M., & Kestring, T. (2022, October). Optimizing wireline formation and evaluation operations in deep horizontal wells: Case study. Retrieved from https://doi.org/10.2118/211651-MS

Franco, N., Corona, M., AlMahroos, M., Kestring, T., & Oliveira, A. (2022, October). Practical guide to successful wireline formation pressure testing across highly deviated and depleted sections. Retrieved from https://doi.org/10.2118/211747-MS

Fryklund, B., & Stark, P. (2020, December). Super basins—New paradigm for oil and gas supply. *AAPG Bulletin, 104*, 2507–2519. doi:https://doi.org/10.1306/09182017314

Gamal, H., Alsaihati, A., & Elkatatny, S. (2022, April). Predicting the rock sonic logs while drilling by random forest and decision tree-based algorithms. *J. Energy Resour. Technol, 144*, 043203. doi:https://doi.org/10.1115/1.4051670

Gamal, H., Alsaihati, A., Elkatatny, S., & Abdulraheem, A. (2021). Sonic logs prediction in real time by using random forest technique., (pp. –). Retrieved from https://onepetro.org/armaigs/proceedings-abstract/IGS21/All-IGS21/473055

Gamal, H., Alsaihati, A., Elkatatny, S., Haidary, S., & Abdulraheem, A. (2021). Rock strength prediction in real-time while drilling employing random forest and functional network techniques. *J. Energy Resour. Technol, 143*, 093004.

Garrouch, A. A., & Al-Sultan, A. A. (2019, May). Exploring the link between the flow zone indicator and key open-hole log measurements: an application of dimensional analysis. *Pet. Geosci, 25*, 219–234,. doi:10.1144/petgeo2018-035.

Ghionna, L., Greco, G., Guzzo, A., & Pontieri, L. (2008). Outlier detection techniques for process mining applications. In A. An, S. Matwin, Z. W. Raś, & D. Ślęzak (Eds.). Springer. doi:10.1007/978-3-540-68123-6_17.

Ghosh, S. (2022, July). A review of basic well log interpretation techniques in highly deviated wells. *J. Pet. Explor. Prod. Technol, 12*, 1889–1906,. doi:10.1007/s13202-021-01437-2.

Giem, G., Sheiretov, T., & Couble, Y. (2018). Wireline tractor for through-tubing intervention in wells with barefoot openhole completions. *SPE/ICoTA Well Intervention Conference and Exhibition, SPE, 10*, 011 003 005. Retrieved from https://onepetro.org/SPECTWI/proceedings-abstract/18CTWI/1-18CTWI/220786

Gill, D., Shomrony, A., & Fligelman, H. (1993). Numerical zonation of log suites and logfacies recognition by multivariate clustering. *AAPG Bulletin, 77*, 1781–1791.

Gordon, J. B., Sanei, H., & Pedersen, P. K. (2022). Predicting hydrogen and oxygen indices (HI, OI) from conventional well logs using a Random Forest machine learning algorithm. *Int. J. Coal Geol, 249*, 103903.

Grover, A. (2018). *Anomaly detection for application log data.* Ph.D. dissertation. Retrieved from https://doi.org/10.31979/etd.znsb-bw4d

Gupta, D., & Shah, M. (2022). A comprehensive study on artificial intelligence in oil and gas sector. *Environ. Sci. Pollut. Res, 29*, 50984–50997. doi:https://doi.org/10.1007/s11356-021-15379-z

Habeeb, A. (2023, February). Introduction and investigation into oil well logging operations. *Journal of Engineering, 29*, 72–91. doi:https://doi.org/10.31026/j.eng.2023.02.05

Handhali, S. A., Aamri, M. A., & Sundararajan, N. (2023, November). Beyond conventional approach: hybrid supervised learning and feature selection algorithms for prediction sonic logs - a study in a tight gas sand, North of Oman. *International Journal of Oil, Gas and Coal Technology, 34*, 359–385. doi:https://doi.org/10.1504/IJOGCT.2023.135055

Hanson, M., Hyson, L., Munro, H., & Wheater, G. (2024, May). Wireline cable dynamics & wellbore diagnostics in the deepwater logging environment., *10*, pp. SPWLA-2024-0143. doi:https://doi.org/10.30632/SPWLA-2024-0143

Harter, P. (2019, February). A randomized trial of lymphadenectomy in patients with advanced ovarian neoplasms. *The New England Journal of Medicine, 380*, 822–832. doi:10.1056/NEJMoa1808424.

Haustveit, K. (2020, January). Monitoring the pulse of a well through sealed wellbore pressure monitoring, a breakthrough diagnostic with a multi-basin case study. Retrieved from https://doi.org/10.2118/199731-MS

He, S., Zhu, J., He, P., & Lyu, M. R. (2016, October). Experience report: System log analysis for anomaly detection. (pp. 207–218). IEEE. Retrieved from 10.1109/ISSRE.2016.21

Hiba, M., Ibrahim, A. F., Elkatatny, S., & Ali, A. (2022, January). Prediction of cohesion and friction angle from well-logging data using decision tree and random forest. *Arab. J. Geosci, 15*, 26,. doi:10.1007/s12517-021-09154-0.

Hinkka, M., Lehto, T., Heljanko, K., & Jung, A. (2018, January). Structural feature selection for event logs. In E. Teniente, & M. Weidlich (Ed.). *308*, pp. 20–35. Springer International Publishing. doi:10.1007/978-3-319-74030-0_2.

Hodge, V., & Austin, J. (2004, October). A survey of outlier detection methodologies. *Artif. Intell. Rev, 22*, 85–126,. doi:10.1023/B:AIRE.0000045502.10941.a9.

Howard, J., Tracey, J., Shen, M., & Zhang, S. (2021, May). An AI-Based approach to enhanced fracture resolution in image logs. SPWLA. doi:10.30632/SPWLA-2021-0081

Ibrahim, A. F., Abdelaal, A., & Elkatatny, S. (2022, September). Formation resistivity prediction using decision tree and random forest. *Arab. J. Sci. Eng, 47*, 12183–12191,. doi:10.1007/s13369-022-06900-8.

Inoue, H. (2023). Shear log estimation by rock physics assist machine learning approach an offshore abu dhabi case study. *Shear log estimation by rock physics assist machine learning approach an offshore abu dhabi case study*. Retrieved from https://www.searchanddiscovery.com/abstracts/pdf/2023/91204meosgeo/abstracts/ndx_inoue.pdf

Jaya, M. S., Sharif, A., Abdulkarim, A. A., Riza, G. A., Hajian, M. A., & Emad, E. (2021, December). Accurate pseudo log prediction using machine learning based automatic log regularization and feature augmentation method. doi:10.2118/207230-MS

Jerram, D. A. (2019). Understanding volcanic facies in the subsurface: a combined core, wireline logging and image log data set from the PTA2 and KMA1 boreholes, Big

Island, Hawaii. *Scientific Drilling, 25*, 15–33. doi:https://doi.org/10.5194/sd-25-15-2019

Jha, H. S., Khanal, A., Seikh, H. M., & Lee, W. J. (2022, September). A comparative study on outlier detection techniques for noisy production data from unconventional shale reservoirs. *Journal of Natural Gas Science and Engineering, 105*, 104720. doi:https://doi.org/10.1016/j.jngse.2022.104720

Jia, B., Tsau, J.-S., & Barati, R. (2019, January). A review of the current progress of CO2 injection EOR and carbon storage in shale oil reservoirs. *Fuel, 236*, 404–427. doi:https://doi.org/10.1016/j.fuel.2018.08.103

Jia, W., Zong, Z., Qin, D., & Lan, T. (2023, November). Integrated well-log data and seismic inversion results for prediction of hydrocarbon source rock distribution in W segment, Pearl River Mouth Basin, China. *Geoenergy Sci. Eng, 230*, 212233,. doi:10.1016/j.geoen.2023.212233.

Kato, K. (2020, November). Logarithmic structures of fontaine-illusie. II - Logarithmic flat topology. *Tokyo Journal of Mathematics, 10*, 1-29. doi:https://doi.org/10.3836/tjm/1502179316

Kausik, R., Fellah, K., Rylander, E., Singer, P. M., Lewis, R. E., & Sinclair, S. M. (2016, August). NMR relaxometry in shale and implications for logging. *Petrophysics, 57*, 339–350.

Kelkouli, M., Arab, H. H., Amor, N. B., Kecili, K., & Mokhtari, N. (2019, September). Multidisciplinary approach for unconventional reservoirs characterisation by integrating wireline openhole logging Techniques–Electric and sonic to formation testing. SPE. Retrieved from https://doi.org/10.2118/196689-MS

Kim, D., Choi, J., Kim, D., & Byun, J. (2020, December). Predicting mineralogy by integrating core and well log data using a deep neural network. *J. Pet. Sci. Eng, 195*, 107838,. doi:10.1016/j.petrol.2020.107838.

Kim, Y., Hardisty, R., Torres, E., & Marfurt, K. J. (2018, October). Seismic facies classification using random forest algorithm. SEG. Retrieved from https://doi.org/10.1190/segam2018-2998553.1

Klitten, K., & Agyekum, W. A. (2021). Geophysical wire-line logging of boreholes in nasia river catchment basin. *Geophysical wire-line logging of boreholes in nasia river catchment basin, 10*. Geological Survey of Denmark and Greenland. Retrieved from https://data.geus.dk/pure-pdf/GEUS-R_2021_51_web.pdf

Koroteev, D., & Tekic, Z. (2021, March). Artificial intelligence in oil and gas upstream: Trends, challenges, and scenarios for the future. *Energy AI, 3*, 100041,. doi:10.1016/j.egyai.2020.100041.

Kulananpakdee, K. (2018, March). Through-drill pipe logging and formation sampling in deviated and risky wells. Retrieved from https://doi.org/10.4043/28604-MS

Kumar, B., Khare, S. K., Choudhury, T., & Devaraju, J. (2024, August). A comprehensive literature review for predicting petrophysical well logs using AI/ML techniques. *Nanotechnology Perceptions, 20*, 1019–1034. doi:https://doi.org/10.62441/nano-ntp.v20is6.81

Kumar, P., Nagarajan, A., & Uchil, P. D. (2018). Analysis of cell viability by the MTT assay. *Analysis of cell viability by the MTT assay*. Cold Spring Harbor Protocols. Retrieved from https://cshprotocols.cshlp.org/content/2018/6/pdb.prot095505.short

Kuna, H. D., García-Martinez, R., & Villatoro, F. R. (2014, August). Outlier detection in audit logs for application systems. *Information Systems, 44*, 22–33. doi:https://doi.org/10.1016/j.is.2014.03.001

Lai, J. (2017, June). Fracture detection in oil-based drilling mud using a combination of borehole image and sonic logs. *Marine and Petroleum Geology, 84*, 195–214.

Lai, J. (2022, April). Geophysical well-log evaluation in the era of unconventional hydrocarbon resources: A review on current status and prospects. *Surveys in Geophysics , 43*, 913–957. doi:10.1007/s10712-022-09705-4.

Landauer, M., Onder, S., Skopik, F., & Wurzenberger, M. (2023, June). Deep learning for anomaly detection in log data: A survey. *Machine Learning with Applications, 12*, 100470. doi:https://doi.org/10.1016/j.mlwa.2023.100470

Lee, J., Byun, J., Kim, B., & Yoo, D.-G. (2017, October). Delineation of gas hydrate reservoirs in the Ulleung Basin using unsupervised multi-attribute clustering without well log data. *Journal of Natural Gas Science and Engineering, 46*, 326–337. doi:https://doi.org/10.1016/j.jngse.2017.08.007

Li, H., Yu, H., Cao, N., Tian, H., & Cheng, S. (2021, May). Applications of artificial intelligence in oil and gas development. *Arch. Comput. Methods Eng, 28*, 937–949,. doi:10.1007/s11831-020-09402-8.

Li, J. (2018, March). The first offshore natural gas hydrate production test in South China Sea. *China Geology, 1*, 5–16. doi:https://doi.org/10.31035/cg2018003

Lima, R. P., Suriamin, F., Marfurt, K. J., & Pranter, M. J. (2019, May). Convolutional neural networks as aid in core lithofacies classification. *Interpretation, 7*, 27– 40,. doi:10.1190/INT-2018-0245.1.

Liu, H. (2017). *Principles and Applications of Well Logging.* Springer Berlin Heidelberg. doi:10.1007/978-3-662-53383-3.

Liu, Y. (2019, October). Petrophysical static rock typing for carbonate reservoirs based on mercury injection capillary pressure curves using principal component analysis. *Journal of Petroleum Science and Engineering, 181*, 106175. doi:https://doi.org/10.1016/j.petrol.2019.06.039

Longde, S. U. (2021, June). An analysis of major scientific problems and research paths of Gulong shale oil in Daqing Oilfield, NE China. *Petroleum Exploration and Development, 48*, 527–540. doi:https://doi.org/10.1016/S1876-3804(21)60043-5

Luo, X., Sun, J., Zhang, J., & Liu, W. (2024, June). A new method based on multiresolution graph-based clustering for lithofacies analysis of well logging. *Comput. Geosci, 28*, 491–502,. doi:10.1007/s10596-024-10277-y.

Luthi, S. M. (2001). *Geological well logs: Their use in reservoir modeling.* Springer-Verlag Berlin Heidelberg New York. Retrieved from https://books.google.com/books?hl=en&lr=&id=mEp1SmQqiWQC&oi=fnd&pg=PA3&dq=well+log&ots=W9Av1oDQLB&sig=C3_PPebjXZKYStf8jX8izFzX2BU

Ma, Y. Z. (2011, May). Lithofacies clustering using principal component analysis and neural network: Applications to wireline logs. *Math. Geosci, 43*, 401–419,. doi:10.1007/s11004-011-9335-8.

Mahmoud, M., Elkatatny, S., Ramadan, E., & Abdulraheem, A. (2016, October). Development of lithology-based static Young's modulus correlations from log data based on data clustering technique. *Journal of Petroleum Science and Engineering, 146*, 10–20. doi:https://doi.org/10.1016/j.petrol.2016.04.011

Manakhayev, R., Tussupbayev, I., Sargunanov, M., Kalzhekov, N., Kulov, G., & Tlepbergenov, N. (2020, October). Use of logging while drilling equipment in washdown mode under managed pressure drilling conditions in tengiz. Retrieved from https://doi.org/10.2118/202518-MS

Masucci, M., Brusoni, S., & Cennamo, C. (2020, February). Removing bottlenecks in business ecosystems: The strategic role of outbound open innovation. *Research Policy, 49*, 103823. doi:https://doi.org/10.1016/j.respol.2019.103823

Matinkia, M., Hashami, R., Mehrad, M., Hajsaeedi, M. R., & Velayati, A. (2023, March). Prediction of permeability from well logs using a new hybrid machine learning algorithm. *Petroleum, 9*, 108–123. doi:https://doi.org/10.1016/j.petlm.2022.03.003

Miah, M. I., Ahmed, S., Zendehboudi, S., & Butt, S. (2020, October). Machine learning approach to model rock strength: Prediction and variable selection with aid of log data. *Rock Mech. Rock Eng, 53*, 4691–4715,. doi:10.1007/s00603-020-02184-2.

Mishra, A. K. (2020, August). Latest wireline conveyance technologies set a new world record, achieving gravity descent to 79 degrees in open hole. SPWLA.

Mishra, D., Singh, K., Mishra, S., Gupta, S., Kumar, A., & Agrawal, G. (2019, April). Don't let water production hinder your ability to meet production targets-a comprehensive wireline logging guide to rejuvenate fields with high water cuts. SPE. Retrieved from https://doi.org/10.2118/194557-MS

Misra, S., Osogba, O., & Powers, M. (2019). Unsupervised outlier detection techniques for well logs and geophysical data. *Mach Learn Subsurf. Charact, 1*.

Mo'men, A., Darwish, M., Abdelhady, A., & Essa, M. A. (2017). Structural and Lithostratigraphic evolution of Al Baraka Oil field, Komombo Basin, Upper Egypt as deduce from 2D seismic lines and well logging data. *Journal of Basic and Environmental Sciences, 4*, 149–169.

Mohammadpoor, M., & Torabi, F. (2020, December). Big Data analytics in oil and gas industry: An emerging trend. *Petroleum, 6*, 321–328. doi:https://doi.org/10.1016/j.petlm.2018.11.001

Moussa, T., Elkatatny, S., Mahmoud, M., & Abdulraheem, A. (2018, March). Development of new permeability formulation from well log data using artificial intelligence approaches. *J. Energy Resour. Technol, 140*. doi:10.1115/1.4039270.

Ngo, H. (2018, August). Rig time optimization and risk mitigation utilizing through drillpipe logging technique with slim wireline logging suite. SPE. Retrieved from https://doi.org/10.2118/191015-MS

Nguyen, H., Savary-Sismondini, B., Patacz, V., Jenssen, A., Kifle, R., & Bertrand, A. (2022). Application of random forest algorithm to predict lithofacies from well and seismic data in Balder field, Norwegian North Sea. *AAPG Bulletin, 106*, 2239–2257. doi:https://doi.org/10.1306/08142221032

Nguyen, T., Gosine, R. G., & Warrian, P. (2020). A systematic review of big data analytics for oil and gas industry 4.0. *IEEE Access, 8*, 61183–61201. doi:10.1109/ACCESS.2020.2979678

Nichol, A. Q., & Dhariwal, P. (2021). Improved denoising diffusion probabilistic models. *International conference on machine learning, PMLR, 139*, 8162–8171. Retrieved from https://proceedings.mlr.press/v139/nichol21a.html

Nordin, T. B., Azid, A. A., Ong, L. W., & Abdullah, M. Z. (2024, February). Advanced wireline log acquisition and innovative data processing, along with an integrated petrophysical methodology, unleash the full potential of cycle II/III reservoir offshore sarawak. Retrieved February 12, 2024, from https://doi.org/10.2523/IPTC-24313-MS

Nyakilla, E. E. (2024, September). Evaluation of reservoir porosity and permeability from well log data based on an ensemble approach: A comprehensive study incorporating experimental, simulation, and fieldwork data. *Natural Resources Research , 34*, 383-408. doi:10.1007/s11053-024-10402-9.

Oliveira, F. V. (2022). Igneability feature: an effective, easy and low-cost way to identify basic igneous rocks using wireline well logs in open hole wells. *Braz. J. Geophys, 40*, 19–41.

Oliveira, L. A., & Carneiro, C. (2021). Synthetic geochemical well logs generation using ensemble machine learning techniques for the Brazilian pre-salt reservoirs. *Journal of Petroleum Science and Engineering, 196*, 108080. doi:https://doi.org/10.1016/j.petrol.2020.108080

Otchere, D. A., Ganat, T. O., Gholami, R., & Ridha, S. (2021, May). Application of supervised machine learning paradigms in the prediction of petroleum reservoir properties: Comparative analysis of ANN and SVM models. *J. Pet. Sci. Eng, 200*, 108182,. doi:10.1016/j.petrol.2020.108182.

Otchere, D. A., Ganat, T. O., Ojero, J. O., Tackie-Otoo, B. N., & Taki, M. Y. (2022, January). Application of gradient boosting regression model for the evaluation of feature selection techniques in improving reservoir characterisation predictions. *Journal of Petroleum Science and Engineering, 208*, 109244. doi:https://doi.org/10.1016/j.petrol.2021.109244

Partouche, A. (2020, June). A technology update in wireline formation testing. SPWLA. Retrieved from https://doi.org/10.30632/SPWLA-5035

Patterson, L. A. (2017, March). Unconventional oil and gas spills: Risks, mitigation priorities, and state reporting requirements. *Environ. Sci. Technol, 51*, 2563–2573,. doi:10.1021/acs.est.6b05749.

Peterson, R. A., & Cavanaugh, J. E. (2020, November). Ordered quantile normalization: a semiparametric transformation built for the cross-validation era. *J. Appl. Stat, 47*, 2312–2327,. doi:10.1080/02664763.2019.1630372.

Pootisirakorn, M. (2018). *Failure prediction in open-hole wireline logging of oil and gas drilling operation using support vector machine.* Ph.D. dissertation. Retrieved from https://digital.car.chula.ac.th/chulaetd/2289/

Popa, A. S., Cassidy, S., & Jikich, S. (2016, June). Clustering-based optimal perforation design using well logs. *J. Sustain. Energy Eng, 3*, 318–332,. doi:10.7569/JSEE.2016.629503.

Predicting mineralogical composition in unconventional formations using machine learning and well logging data | IPTC international petroleum technology conference | OnePetro. (2024, September). *Predicting mineralogical composition in unconventional formations using machine learning and well logging data | IPTC international petroleum technology conference | OnePetro*. Retrieved from https://onepetro.org/IPTCONF/proceedings-abstract/24IPTC/All-24IPTC/542532

Projecting petrophysical logs at the bit through multi-well data analysis with machine learning | SPE offshore europe conference and exhibition | OnePetro. (2024, September). *Projecting petrophysical logs at the bit through multi-well data analysis with machine learning | SPE offshore europe conference and exhibition | OnePetro*. Retrieved from https://onepetro.org/SPEOE/proceedings-abstract/23OE/3-23OE/530194

Puskarczyk, E. (2020). Application of multivariate statistical methods and artificial neural network for facies analysis from well logs data: An example of Miocene deposits. *Energies, 13*, 1548. doi:https://doi.org/10.3390/en13071548

Qiao, X., Shao, Z., Bao, W., & Rong, Q. (2017). Fiber Bragg grating sensors for the oil industry. *Sensors, 17*, 429. doi:https://doi.org/10.3390/s17030429

Quintero-Bermudez, R. (2018, September). Compositional and orientational control in metal halide perovskites of reduced dimensionality. *Nature Mater, 17*, 900–907. doi:https://doi.org/10.1038/s41563-018-0154-x

Radwan, A. E. (2021, June). Modeling pore pressure and fracture pressure using integrated well logging, drilling based interpretations and reservoir data in the Giant El Morgan oil Field, Gulf of Suez, Egypt. *Journal of African Earth Sciences, 178*, 104165. doi:https://doi.org/10.1016/j.jafrearsci.2021.104165

Rahaman, M. S., Vasant, P. M., Jufar, S. R., & Watada, J. (2020). Feature selection-based artificial intelligence techniques for estimating total organic carbon from well logs. (p. 1529). IOP Publishing. Retrieved from https://iopscience.iop.org/article/10.1088/1742-6596/1529/4/042084/meta

Rajaee, T., Ebrahimi, H., & Nourani, V. (Eds.). (2019, May). A review of the artificial intelligence methods in groundwater level modeling . *Journal of Hydrology, 572*, 336-351. doi:https://doi.org/10.1016/j.jhydrol.2018.12.037

Rebelle, M., & Lalanne, B. (2014, November). Rock-typing in carbonates: A critical review of clustering methods. (p. 041 070 002). SPE. Retrieved from https://doi.org/10.2118/171759-MS

Riadi, I., Istiyanto, J. E., Ashari, A., & Subanar. (2013, June). Log analysis techniques using clustering in network forensics. *Log analysis techniques using clustering in network forensics*. Retrieved from http://arxiv.org/abs/1307.0072

Rossi, T. B., Martinez, A., & Useche, M. A. (2023, October). DAS VSP using hybrid wireline logging cable in deepwater wells. *OTC-32904-MS.* OTC. Retrieved from https://doi.org/10.4043/32904-MS

Rostamian, A., Heidaryan, E., & Ostadhassan, M. (2022, January). Evaluation of different machine learning frameworks to predict CNL-FDC-PEF logs via hyperparameters optimization and feature selection. *Journal of Petroleum Science and Engineering, 208*, 109463. doi:https://doi.org/10.1016/j.petrol.2021.109463

Rui, Z. (2017, June). A quantitative oil and gas reservoir evaluation system for development. *Journal of Natural Gas Science and Engineering, 42*, 31–39. doi:https://doi.org/10.1016/j.jngse.2017.02.026

Sales, I. M. (2019, April). Wireline logging in managed pressure drilling well on a DP rig. SPE. Retrieved from https://doi.org/10.2118/194542-MS

Sani, M. F., Zelst, S. J., & Aalst, W. M. (2018, June). Repairing outlier behaviour in event logs. In W. Abramowicz, & A. Paschke (Ed.). *320*, pp. 115–131. Springer International Publishing. doi:10.1007/978-3-319-93931-5_9.

Sebtosheikh, M. A., Motafakkerfard, R., Riahi, M. A., Moradi, S., & Sabety, N. (2015, March). Support vector machine method, a new technique for lithology prediction in an Iranian heterogeneous carbonate reservoir using petrophysical well logs. *Carbonates Evaporites, 30*, 59–68,. doi:10.1007/s13146-014-0199-0.

Serra, O. E. (1983). *Fundamentals of well-log interpretation.* Retrieved from https://www.osti.gov/biblio/5373704/1000

Sharma, A., Burak, T., & Nygaard, R. (2023). Projecting Petrophysical Logs at the Bit through Multi-Well Data Analysis with Machine Learning. *SPE Offshore Europe Conference & Exhibition, Aberdeen, Scotland, UK*, (pp. SPE-215587-MS).

Shrivastava, C., Zhang, T., Yang, S., Forest, N.-B., & Nir, I. L. (2021). Real-time geology: Review of recent advances made with logging while drilling images. Retrieved from https://onepetro.org/OMCONF/proceedings-abstract/OMC21/All-OMC21/473162

Shuangfang, L. (2018). Classification of microscopic pore-throats and the grading evaluation on shale oil reservoirs. *Petroleum Exploration and Development, 45*, 452–460. doi:https://doi.org/10.1016/S1876-3804(18)30050-8

Silveira, J. R., Figueiredo, J. J., Lima, C. R., Gonçalvez, J. F., & Amaral, M. L. (2024, October). S-wave log construction through semi-supervised regression clustering using machine learning: A case study of North Sea fields. *Journal of Applied Geophysics, 229*, 105476. doi:https://doi.org/10.1016/j.jappgeo.2024.105476

Singh, H., Seol, Y., & Myshakin, E. M. (2020, October). Automated well-log processing and lithology classification by identifying optimal features through unsupervised and supervised machine-learning algorithms. *SPE J, 25*, 2778–2800,. doi:10.2118/202477-PA.

Singh, M. (2020, November). Machine learning assisted petrophysical logs quality control, editing and reconstruction. SPE. Retrieved from https://doi.org/10.2118/202977-MS

Sircar, A., Yadav, K., Rayavarapu, K., Bist, N., & Oza, H. (2021, December). Application of machine learning and artificial intelligence in oil and gas industry. *Petroleum Research, 6*, 379–391. doi:https://doi.org/10.1016/j.ptlrs.2021.05.009

Siregar, R. (2024, February). Record-breaking on wireline data acquisition: Leveraging large dimension of wireline formation tester in slim hole and high deviation Well–Pushing the boundaries of wellbore clearance issue. OTC. Retrieved from https://doi.org/10.4043/34918-MS

Siregar, R., Ramli, S., Turner, D., Prachayarittikool, K., & Iffwad, M. (2022, March). Mitigation of slim open hole wireline logging risk under high overbalance conditions. *OTC-31373-MS.* OTC. Retrieved from https://doi.org/10.4043/31373-MS

Siwach, D. S. (2022). Anomaly detection for web log data analysis: A review. *Journal of Algebric Stastics, 13*, 129–148.

Soeder, D. J. (2018). The successful development of gas and oil resources from shales in North America. *Journal of Petroleum Science and Engineering, 163*, 399–420. doi:https://doi.org/10.1016/j.petrol.2017.12.084

Solanki, P., Baldaniya, D., Jogani, D., Chaudhary, B., Shah, M., & Kshirsagar, A. (2022, March). Artificial intelligence: New age of transformation in petroleum upstream. *Petroleum Research, 7*, 106–114. doi:https://doi.org/10.1016/j.ptlrs.2021.07.002

Song, T. (2023, December). A novel well-logging data generation model integrated with random forests and adaptive domain clustering algorithms. *Geoenergy Science and Engineering, 231*, 212381. doi:https://doi.org/10.1016/j.geoen.2023.212381

Sreevidya, S. S. (2014). A survey on outlier detection methods. *International Journal of Computer Science and Information Technologies, 5*, 8153-8156. Retrieved from https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=759d21394435103ca62d23955f9bb99edda5e878

Srivardhan, V. (2022, September). Adaptive boosting of random forest algorithm for automatic petrophysical interpretation of well logs. *Acta Geod. Geophys, 57*, 495–508,. doi:10.1007/s40328-022-00385-5.

Stuov, N., Zaytsev, A., Ismailova, L., Kovalev, D., Katterbauer, K., & Marsala, A. (2022, February). Out of distribution detection of well logs for AI-Assisted formation evaluation. IPTC. Retrieved from https://doi.org/10.2523/IPTC-22097-MS

Sun, Y., Zhang, J., & Zhang, Y. (2024). Adaboost algorithm combined multiple random forest models (Adaboost-RF) is employed for fluid prediction using well logging data. *Physics of Fluids, 36*. doi:https://doi.org/10.1063/5.0179422

Suri, N. N., M, N. M., & Athithan, G. (2019). A Survey on Anomaly Detection Strategies. Springer International Publishing. doi:10.1007/978-3-030-05127-3.

Szabó, N. P., Kilik, R., & Dobróka, M. (2023, May). Robust reservoir identification by multi-well cluster analysis of wireline logging data. *Heliyon, 9*, e15957. Retrieved from https://www.cell.com/heliyon/fulltext/S2405-8440(23)03164-X

Tabasi, S. (2022). Optimized machine learning models for natural fractures prediction using conventional well logs. *Fuel, 326*, 124952. doi:https://doi.org/10.1016/j.fuel.2022.124952

Talebkeikhah, M., Sadeghtabaghi, Z., & Shabani, M. (2021). A comparison of machine learning approaches for prediction of permeability using well log data in the hydrocarbon reservoirs. *Journal of Human, Earth, and Future, 2*, 82–99. doi:http://dx.doi.org/10. 28991/HEF-2021-02-02-01

Talreja, R. (2023, April). Well log data preparation and effective utilization of drilling parameters using data science based approaches. In Y. B. Wah, M. W. Berry, A. Mohamed, & D. Al-Jumeily (Ed.). *165*, pp. 388–402. Springer Nature Singapore. doi:10.1007/978-981-99-0741-0_28.

Tavakkoli, V., & Amini, A. (2006). Application of multivariate cluster analysis in logfacies determination and reservoir zonation, case study of Marun Field, South of Iran.

Accessed: Oct, 32, 69-75. Retrieved from https://www.sid.ir/EN/VEWSSID/J_pdf/80420060203.pdf

Tawfeeq, Y. J., Najmuldeen, M. Y., & Ali, G. H. (2020). Optimal statistical method to predict subsurface formation permeability depending on open hole wireline logging data: A comparative study. *Periodicals of Engineering and Natural Sciences (PEN), 8*, 736–749.

Tiab, D., & Donaldson, E. C. (2024). *Petrophysics: theory and practice of measuring reservoir rock and fluid transport properties.* Elsevier. Retrieved from https://books.google.com/books?hl=en&lr=&id=WOTPEAAAQBAJ&oi=fnd&pg=PP1&dq=well+LOG+&ots=W9YRI6GmvH&sig=-ISzgT1VnF6iwBlfbA_NilaDLvE

Tjioe, M., McCaslin, J., Brege, J. J., Oyler, K. W., Rodrigue, K., & Adeyemi, W. A. (2022, October). Wireline logging in high overbalance high-temperature high-pressure deepwater gulf of mexico well with water-based reservoir drilling fluid. SPE. Retrieved from https://doi.org/10.2118/210021-MS

Torghabeh, A., Rezaee, R., Moussavi-Harami, R., Pradhan, B., Kamali, M., & Kadkhodaie-Ilkhchi, A. (2014, January). Electrofacies in gas shale from well log data via cluster analysis: A case study of the Perth Basin, Western Australia. *Open Geosci, 6*. doi:10.2478/s13533-012-0177-9.

Tripoppoom, S. (2019, July). Application of assisted history matching workflow to shale gas well using EDFM and neural network-markov chain monte carlo algorithm . In W. Yu, K. Sepehrnoori, & J. Miao (Ed.). Retrieved from https://doi.org/10.15530/urtec-2019-659

Urang, J. G., Ebong, E. D., Akpan, A. E., & Akaerue, E. I. (2020, December). 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. *J. Appl. Geophys, 183*, 104207,. doi:10.1016/j.jappgeo.2020.104207.

Wang, H., Bah, M. J., & Hammad, M. (2019, August). Progress in outlier detection techniques: A survey. *IEEE Access, 7*, 107964–108000. doi:10.1109/ACCESS.2019.2932769

Wang, J., Cao, J., & Yuan, S. (2020, November). Shear wave velocity prediction based on adaptive particle swarm optimization optimized recurrent neural network. *J. Pet. Sci. Eng, 194*, 107466,. doi:10.1016/j.petrol.2020.107466.

Wheater, G., Huyton, S., Miller, L., Gerrity, R., & Ford, R. (2018, November). Wireline cable protection: Enabling fluid sampling in high-risk wellbores. Retrieved from https://doi.org/10.2118/193232-MS

Wu, P.-Y., Jain, V., Kulkarni, M. S., & Abubakar, A. (2018, October). Machine learning-based method for automated well-log processing and interpretation. SEG. Retrieved from https://doi.org/10.1190/segam2018-2996973.1

Wurzenberger, M., Skopik, F., Fiedler, R., & Kastner, W. (2017, June). Applying high-performance bioinformatics tools for outlier detection in log data. (pp. 1–8). IEEE. Retrieved from https://ieeexplore.ieee.org/abstract/document/7985760/

Yang, H., Pan, H., Ma, H., Konaté, A. A., Yao, J., & Guo, B. (2016). Performance of the synergetic wavelet transform and modified K-means clustering in lithology classification using nuclear log. *Journal of Petroleum Science and Engineering, 144*, 1–9. doi:https://doi.org/10.1016/j.petrol.2016.02.031

Ye, Y., Jiang, Z., Liu, X., Wang, Z., & Gu, Y. (2024, September). Logging identification method for reservoir facies in fractured-vuggy dolomite reservoirs based on AI: A case study of ediacaran dengying formation, sichuan basin, china. *Applied Sciences, 14*, 7504. doi:https://doi.org/10.3390/app14177504

Yehia, T., Khattab, H., Tantawy, M., & Mahgoub, I. (2022). Improving the shale gas production data using the angular-based outlier detector machine learning algorithm. *Journal of University of Shanghai for Science and Technology, 24*, 152–172.

Zakaria, K. M., Nawawi, A., & Salin, A. S. (2016). Internal controls and fraud–empirical evidence from oil and gas company. *Journal of Financial Crime, 23*, 1154–1168. doi:https://doi.org/10.1108/JFC-04-2016-0021

Zhang, Y., Meratnia, N., & Havinga, P. (2010). Outlier detection techniques for wireless sensor networks: A survey. *IEEE Communications Surveys & Tutorials , 12*, 159–170. doi:10.1109/SURV.2010.021510.00088

Zhao, L. (2021, July). Fluid and lithofacies prediction based on integration of well-log data and seismic inversion: A machine-learning approach. *GEOPHYSICS, 86*, 151– 165,. doi:10.1190/geo2020-0521.1.

Zhou, K. (2020, January). Fast prediction of reservoir permeability based on embedded feature selection and LightGBM using direct logging data. *Measurement Science and Technology, 31*, 045101. doi:10.1088/1361-6501/ab4a45

Zhu, G., Chen, F., Wang, M., Zhang, Z., Ren, R., & Wu, L. (2018). Discovery of the lower Cambrian high-quality source rocks and deep oil and gas exploration potential in the Tarim Basin, China. *AAPG Bulletin, 102*, 2123–2151. doi:https://doi.org/10.1306/03141817183

Zhu, L., Zhou, X., & Zhang, C. (2021, December). Rapid identification of high-quality marine shale gas reservoirs based on the oversampling method and random forest algorithm. *Artificial Intelligence in Geosciences, 2*, 76–81. doi:https://doi.org/10.1016/j.aiig.2021.12.001

APPENDIX- 1
LIST OF PUBLICATIONS

Name: Bishnu Kumar

Orcid ID: 0009-0009-5319-122X

- Kumar, B., Khare, S. K., Choudhury, T., & Devaraju J. (2024). A Methodology for missing Well Logs Prediction using AI/ML Models – A Case Study from Volve Oilfield, Norway. *Nanotechnology Perceptions*, *20*(S10). https://doi.org/10.62441/nano-ntp.v20is10.51

- Kumar, B., Khare, S. K., Choudhury, T., & Devaraju J. (2024). A Comprehensive literature review for predicting petrophysical well logs using AI/ML techniques. *Nanotechnology Perceptions*, *20*(S6). https://doi.org/10.62441/nano-ntp.v20is6.81

APPENDIX- 2

PLAGIARISM REPORT

Bishnu Kumar

Integrated Approach on AI/ML assisted Log reconstruction methodology for estimation of the missing/non-recorded logs

Quick Submit

Quick Submit

University of Petroleum & Energy Studies

Document Details

Submission ID
trn:oid:::1:3263371181

Submission Date
May 29, 2025, 4:59 PM GMT+5:30

Download Date
May 30, 2025, 11:13 AM GMT+5:30

File Name
Bishnu_Kumar_PhD_Thesis_29_May.pdf

File Size
4.1 MB

210 Pages

69,787 Words

391,754 Characters

turnitin Page 1 of 215 - Cover Page Submission ID trn:oid:::1:3263371181

1% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text
- Small Matches (less than 14 words)

Match Groups

- 35 Not Cited or Quoted 1%
  Matches with neither in-text citation nor quotation marks
- 3 Missing Quotations 0%
  Matches that are still very similar to source material
- 0 Missing Citation 0%
  Matches that have quotation marks, but no in-text citation
- 0 Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

1% Internet sources
1% Publications
1% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

turnitin Page 2 of 215 - Integrity Overview Submission ID trnoid:::1:3263371181

UPES
UNIVERSITY WITH A PURPOSE

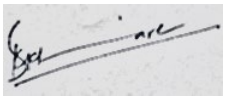
PLAGIARISM CERTIFICATE

1. We Dr. Sunil Kumar Khare (Internal Guide), Dr. Tanupriya Choudhury (Co Guide/ External Guide) certify that the Thesis titled

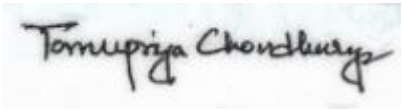
Integrated approach on AI/ML assisted log reconstruction methodology for estimation of the missing/non-recorded logs.

submitted by Scholar Mr/ Ms Bishnu Kumar having SAP ID 500087073 has been run through a Plagiarism Check Software and the Plagiarism Percentage is reported to be 1 %.

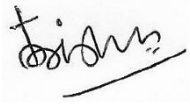
2. Plagiarism Report generated by the Plagiarism Software is attached.



Signature of the Internal Guide



Signature of External Guide/ Co Guide



Signature of the Scholar