Integrating Ensemble Learning and Well-Log Analysis for Sonic Log Prediction in Hydrocarbon Reservoirs

Dr. Suraj Arya, Yogesh Malik

[a]*Department of Computer Science and Information Technology, Central University of Haryana, , Mahendergarh, 123031, Haryana, India*

Abstract

This study shows the applications of ensemble learning models for predicting the sonic logs using the healthy log data from Midland Basin (Texas, USA). The study's primary objective is to assess the feasibility of leveraging data from one section of a geological block to predict logs in another section by emphasizing the influence of spatial and geological variability. The research framework also identifies the optimal combination of input logs by revealing that Resistivity (LLD), Density (RHOB), and Neutron Porosity (NPHI) are the most critical log variables for accurate prediction. These logs are closely associated with lithology, porosity, and rock texture, significantly influencing sonic travel time. Among the machine learning models used, ensemble-based approaches, which include Random Forest, XGBoost, and Cubist, emerged as top performers. Random Forest shows the highest predictive accuracy of MAE = 0.42, RMSE = 0.7, and R^2 = 0.9938. These models have effectively captured complex patterns within the dataset, outperforming the simpler Standalone regression models like Linear Regression, Support Vector Regressor, or Partial Least Squares. Additionally, data preprocessing techniques like outlier removal and extensive data cleaning further helped enhance the model's performance by reducing residual errors. Beyond the numerical accuracy, the models provided valuable geological insights. Analysis of the healthy log data across the depths of 5000 to 11,000 ft revealed the lithological transitions from the porous sandstone to compact limestone and the shale layers with sonic wave travel time variations aligning well with geological features. These

Email addresses: surajarya@cuh.ac.in (Dr. Suraj Arya), yogesh777malik@gmail.com (Yogesh Malik)

findings confirm the ability of the ensemble learning model to effectively incorporate geological complexities into their predictions by reinforcing their utility in the subsurface characterization. The outcome of this study carries significant cost-saving implications for the hydrocarbon industry. By utilizing existing well logs data, the proposed ensemble learning framework reduces the necessity for additional field measurements, offering a more cost-effective alternative for reservoir characterization. Predicting logs from the available data optimizes resource allocation, production quality, or efficiency, enhances exploration , and supports timely decision-making in hydrocarbon exploration. To the best of the authors' knowledge, this study is among the first to investigate the significance of ensemble models compared to standalone regression models in predicting compressional sonic logs and in healthy analysis in the Midland Basin, Texas, USA. While similar research may have been conducted in other regions, few studies have specifically examined this approach in the Midland Basin, emphasizing the significance and novelty of this work.

Keywords: well logging, Sonic log, Hydrocarbon Reservoir, Machine Learning, Ensemble Learning, Random Forest, XGBoost

1. Introduction

Well logging which is also know as borehole logging, is seen as a fundamental technique in hydrocarbon (oil and gas) industry used to see and analyze the geological formations that are encountered during the process of drilling. This process involves the deployment of the specialized tools, like **SONDE**, into the well-bore to measure the different rock properties, which includes electrical, acoustic and radioactive characteristics. The data collected is then transmitted to the surface and are analyzed by the experts to interpret the subsurface formations. Well logging primary objective is to gather the comprehensive data from rock formation, which is important for identifying hydrocarbon reservoirs, evaluating the quality of hydrocarbon reservoir and optimizing the production strategies.

These well logs provide detailed descriptions of the subsurface properties of rocks including their type, fluid saturation, porosity, pressure, temperature and permeability. This information is very crucial for oil and gas industries, as it helps them to determine the viability of continuing drilling operations on the basis of the presence of the hydrocarbons. Moreover, well logs are also useful in assessing the potential for the usable water from the well.

Their are various techniques that are employed in the well logging, each serving a specific purpose of their own. Among these techniques, **Sonic Logging**

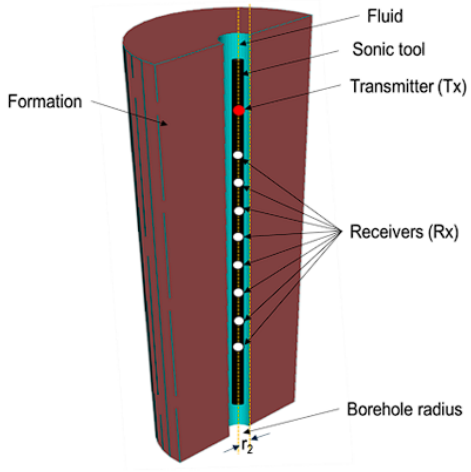

Figure 1: Sonic logging tool

stands out as the most vital method for the evaluation of subsurface rock properties. Sonic logs also referred as the **Acoustic logs** sometimes, measure the time taken by the sound waves to travel through the earth crust and rock formations, this property of wave propagation is known as **slowness**, usually expressed in microseconds per foot ($\mu s/ft$). The sonic logging tool make high frequency sound waves, which are transmitted through the **Transmitter (**T_x**)** and recorded by the **Receivers (**R_x**)** as shown in Figure 1 (source). The travel time of the sound waves can be easily influenced by the rock's density, fluids present in the pore spaces and elastic properties.

When sound energy travels through the rock and reaches back to the receiver, different types of waves arrive at different times due to their varying velocities and pathways. The first which arrive is the P-wave (compressional wave) as seen in the Figure 2 [1] , whose travel time is the fastest but has a relatively low amplitude. Following the P-wave comes S-wave (shear wave), which is slower and has a high amplitude ; although it cannot propagate through fluids. Other wave types, like Rayleigh waves, Stoneley waves and mud waves also travel through the medium but are typically filtered out during analysis. The primary measurement in a sonic log can be seen as the time taken by the P-wave from transmitter to the receiver. This travel time is very crucial for calculating the wave velocity within the formations, providing essential data for subsurface evaluation.

Sonic logs can also be negatively influenced by the factors such as changes in

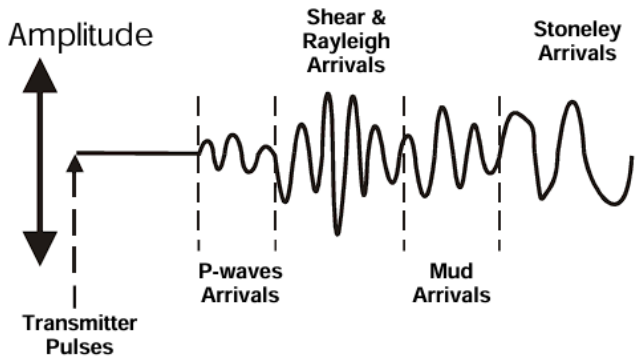

Figure 2: The geophysical wave train received by sonic logging tool

the diameter of the well, mud invasion and damage to the borehole wall. Rough borehole walls can introduce vibrational noise into the data, but it can be minimized by using centralizers and logging in an upward direction. However, some of these issues can be alleviated depending on the design of the tool. For instance, compensated sonic tool have two transmitters that help in correcting the variations in the diameter of the well and borehole damage. Furthermore, there are many tools that are available which examine different parameters. Table 1 below outlines the types of tools and their applications in various scenarios.

The significance of the sonic logs in the oil and gas industry are exceptional hence can't be ignored. They provide seismic calibration, enabling the conversion of the seismic travel time to depth, thereby facilitating accurate mapping of the subsurface structures and accurate geological modeling. The mechanical which are derived from the sonic logs can be vital for ensuring well-bore stability and designing better well completions mainly in hydraulic fracturing operations.

1.1. Problems and Challenges with Sonic logs

A significant problem that is associated with the utilization of the sonic logs in the hydrocarbon sector is **Limited Availability of Data**. Sonic logs are not usually recorded for every well, it can happen often due to operational priorities or specific objectives tied to the drilling project. In case of the older wells, sonic logs may be completely missing because the necessary logging techniques were either unavailable or not widely adopted at that time. This lack of data can pose some really big problems for accurate reservoir characterization and subsurface analysis, which can further lead to the uncertainties in decision - making processes.

Another hurdle with sonic logs is **High expense of obtaining the Sonic log**

Table 1: Applications of Different Wave Components in Sonic Logging

Wave component (velocity and amplitude)	**Application**
Full waveform	velocity profiling, lithology, fracture detection
	Cement-bond evaluation
Compressional wave velocity (Vp)	Porosity and lithology
	Hydrocarbon resource evaluation
	Cement-bond evaluation
Shear-wave velocity (Vs)	Fracture detection and evaluation
	Permeability and gas detection
	Mechanical properties
Stoneley waves	Fracture and permeability evaluation

data. The acquisition of this data needs advanced logging equipments and skilled personnel's how handle the equipments and analyze the data, which makes it a financially demanding undertaking. This is a most true problem for well located in the remote and difficult to access locations. In the explorations projects with limited budgets, operators frequently opt for the more cost-effective logging tools and techniques like gamma-ray or resistivity logs. Additionally, difficulties in the tool deployment, poor borehole conditions like washouts and limitations in the depth of investigations, which can result in less accurate readings in some areas such as gas zones can also produce a challenge. Hence, logging operations in deep or remote well may incur higher costs due to increase in the complexity and associated technical or operational risks.

Despite these risks or challenges, companies may hesitate to acquire sonic logs

because their are difficulties in securing funding for the hydrocarbon projects, particularly in light of the global energy transition that emphasizes a shift from fossil fuels to non-carbon based energy sources.

To address these limitations, there is an continuous demand for the **Reliable Predictive Models** that can predict the sonic log data using the information collected from other, more easily available well logs. Such predictive models not only provide a cost-effective alternative but also helps in bridge gaps in the available datasets, by ensuring that the subsurface evaluations are not compromised by missing data. In cases where the sonic logs are presented , these models can also utilize existing logs, such as gamma-ray, density or resistivity logs to generate more accurate estimation of sonic log parameter. However, the effectiveness of these models is heavily depends on the quality and diversity of the input data.

1.2. Research Objectives of the Study

1. **Develop Generalizable Machine Learning Models**: The primary objective of the research is to develop machine learning models that can predict sonic logs more accurately while maintaining adaptability across different geological formations. Unlike the already existing models that are often used to a region-specific dataset, the proposed models will be useful in handling well logs from various sections of oil wells, focusing mainly on formations within Midland Basin, Texas.

2. **Assess Cross-Well Predictive Capabilities**: This study aims to evaluate the ML model ability that are trained on data from one well to predict sonic logs in different wells, within the same block and across the other blocks too in Midland Basin. By testing model's generalized performance, this research will shows us whether they can be trustworthy applied across distinct geological settings, thereby enhancing their practical scalability and applicability in industry.

3. **Identify Optimal Input Logs for Enhanced Prediction Accuracy**: A key objective of research is to determine the most effective combination of well logs. This research will analyze the significance of different well log parameters, including Potassium (POTA), Density (RHOB), Gamma Ray (GR), Neutron Porosity (NPHI), Resistivity (LLD), Thorium (TH) and Uranium (URAN). By identifying the features that are most informative, the study aim to standardize input log selection and improve model's robustness.

4. **Integrate Advanced Data Preprocessing Techniques**: Given the diffi-

culties posed by the noisy, incomplete and missing well log data, this research will explore and implement the advanced data preprocessing methods. Techniques like data imputation, noise reduction and outlier handling well be used to improve the quality of the input data.

5. **Evaluate the Performance of Machine Learning Algorithms**: This research will conduct a comparative analysis of different machine learning models such as standalone models and ensemble models. Traditional methods like Linear regression, Elastic Net Regression, Support vector Regressor or KNN will be assessed alongside ensemble models like Random Forest Regressor, Cubist Model or XGBoost. This study aims to understand how these models capture the nonlinear relationships in well log data and automate feature extraction, by identifying the most effective approach for achieving high generalization and prediction accuracy.

6. **Optimize Models for Practical and Cost-effective Deployment**: The final objective is going to be to balance prediction accuracy with computational efficiency, ensuring that the developed models are suitable for the real-time applications in oil and gas industry. By optimizing ML models for operational feasibility and cost-effectiveness, this research will support practical deployment in the reservoir characterization and subsurface analysis.

2. Literature Review

Several studies have extensively explored the use of machine learning models for predicting sonic logs, which are critical in evaluating hydrocarbon reservoirs. Sonic logs provide valuable insights into subsurface properties, aiding in accurate reservoir characterization and hydrocarbon assessment. Machine learning has proven to be an effective approach in this domain, as demonstrated by various research efforts.

In recent times, [2] supervised machine learning algorithms for classification and regression problems have gained increasing prevalence in log interpretation. Regression techniques, in particular , have been widely applied to predict physical parameters from the logs. For example, wang utilized a back-propagation neural network to predict the formation permeability from wire-line formation tested data. Similarly, Li and Misra (2019) [3] employed Variational Auto-encoders and Long Short-Term Memory (LSTM) networks to model nuclear magnetic resonance T2 distributions, quantifying pore size distributions in shale reservoirs.

Deep learning techniques have also shown significant promise in log interpretation. Chopra (2022) [4] applied a deep neural network to predict porosity

firm the logs in Volve oilfield, located in the Norwegian North Sea. Additionally, regression-based supervised learning methods have been effectively used to reconstruct missing logs. For instance, Meshalkin (2020) [5] combined multiple algorithms, including AdaBoost, Gradient Boosting and Extra Trees, to construct a rock thermal conductivity prediction model. This model accurately predicted the thermal conductivity curve from well logs, demonstrating the versatility of the ensemble methods.

Focusing on sonic log prediction, one study [6] employed machine learning techniques to predict shear sonic logs using input features such as resistivity, gamma ray, neutron porosity and bulk density logs. This approach, implemented with data from two wells in the Poseidon field, Australia, highlighted the potential of machine learning models to extract meaningful patterns from diverse datasets.

In the Anadarko Basin , Gene Expression Programming (GEP) [7] was used to predict sonic logs from gamma ray and deep resistivity logs. This study underscored the flexibility of machine learning methods in addressing the complex geological relationships. Support vector regression (SVR) was also employed in the Anadarko Basin [8] to address missing sonic log data, revealing that multivariable techniques incorporating several log variables delivered more reliable results compared to single variable models.

Deep learning methods, such as Artificial Neural Networks (ANNs) [9] , have also been utilized for sonic log prediction. Input features such as gamma ray, neutron porosity and density logs were used to train ANNs, showing their ability to generate accurate predictions even with a limited set of input features.

In the Erdos Basin [10] , Kernel Extreme Learning Machine (KELM) was applied to predict sonic logs using gamma ray, density and spontaneous potential logs as inputs. This study, which analyzed data from seven wells, demonstrated KELM's capability in processing complex geological data.

A comprehensive study in the Netherlands [11], [12] evaluated multiple machine learning algorithms, including Random Forest Regression, Support Vector Regression and K-Nearest Neighbors (KNN), for sonic log prediction. The analysis used density, gamma ray and porosity logs as input features, showcasing the adaptability of different machine learning models to solve the same problem of prediction.

Another study in the sedimentary basin of Ghana , West Africa [13],[14], employed supervised machine learning algorithms for sonic log prediction. These findings further reinforced the broad applicability of machine learning models in diverse geological and geographical contexts.

Overall, these studies highlights the versatility and effectiveness of machine

learning techniques in predicting sonic logs using commonly available logs as input features. However, existing machine learning models for the prediction of sonic logs are typically trained on the region-specific datasets, such as the Poseidon field in the Browse Basin (Australia), Netherlands F3 block, Sedimentary Basin in Ghana and the Anadarko Basin (Oklahoma). These models rely heavily on local geological conditions and setups of well-logging, which limits their applicability to different regions. Since the geological variations significantly impact the model's performance, there is a need for the approach that can enhance model adaptability and generalization across diverse geological settings.

Another major challenge in ML-based sonic log prediction is inconsistency presented in the selection of the input logs. Different studies have utilized varying combination of logs like gamma ray, bulk density , neutron porosity, and resistivity logs without a standardized approach. This variability often leads to inconsistent model performance across different formations. Identifying a generalizable and optimal set of input logs would improve the robustness and reliability of models. Additionally, dealing with missing values, outliers and noisy or inconsistent data is crucial for the real-time applications. Techniques like data imputation, noise reduction and data augmentation must be used into the model's workflow to ensure reliability when working with diverse data.

Algorithms with high prediction accuracy like Kernel Extreme Learning Machine (KELM) and Gene Expression Programming (GEP) are very good in making a prediction but are computationally intensive, making them impractical for the cost sensitive operations. There is need for the explore trade-offs between model accuracy and computational efficiency. Simplified ensemble models like Random Forest, Extra-tree Regressor , Cubist Model or XGBoost can be effective alternatives, providing or maintaining sufficient predictive quality.

3. Research Methodology

3.1. Dataset Overview

The dataset used in this study is derived from the website source Midland Basin, located in the western part of the Permian Basin, Texas , USA as shown in 3 (source). This region is recognized globally as one of the most productive oil and gas-producing areas, with vast reserves of hydrocarbon including oil and natural gas. Geographically, it spans Ector, Midland, Martin and Glasscock counties and is a cornerstone of the energy sector, employing advanced extraction techniques such as horizontal drilling and hydraulic fracturing.

Geologically, the Midland Basin is a complex system of sedimentary rock layers deposited over million of years. The primary rock formations in this region in-

Figure 3: Midland Basin

clude shale, sandstone, dolomite and limestone, all of which are hydrocarbon-rich. Notably, key shale formations like **Spraberry, Wolfcamp and Clearfork** serve as major sources of oil production 4 (source) . These formations exhibit diverse physical and chemical properties, such as **porosity, resistivity and permeability**, which are essential for understanding reservoir characteristics and assessing production potential.

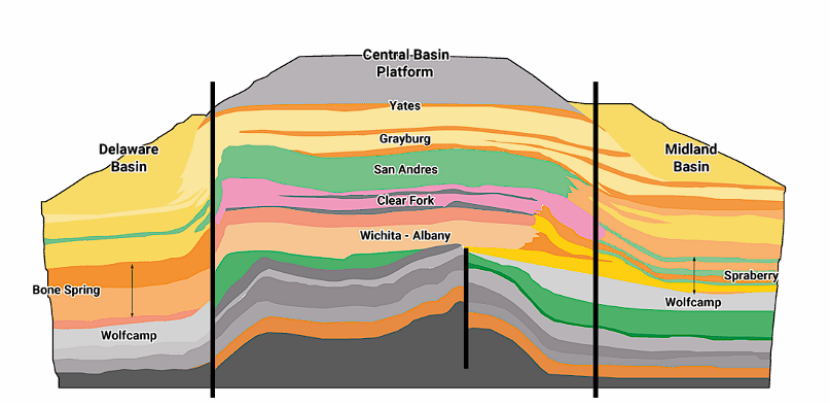

Figure 4: Layers in Midland Basin

The Midland Basin's formations mainly consist of carbonate reef deposits and shallow marine clastic sediments. From youngest to oldest , the formations in-

clude **Tansil, Yates, Seven Rivers, Queen, Grayburg, San Andres, Glorieta, Leonard, Spraberry, Dean and Wolfcamp.**

The study uses raw data extracted from the oil wells in the log files or .las files. These log files have various range of measurements used to determine the subsurface geological properties. Data is then extracted from the log files/.las files and organized into a structured dataset suitable for the analysis. The dataset used is divided into the different training and testing sets. The training set contains **13,020 rows**, while the testing set consist **13,428 rows**. Both of these sets include **29 features/columns**, capturing different petrophysical properties essential for the sonic log prediction.

	DEPT	POTA	URAN	THOR	LLD	GR	CALI	BHVT	RHOB	PE	NPHI	MDT
count	13020.000000	13020.000000	13020.000000	13020.000000	13020.000000	13020.000000	13020.000000	13020.000000	13020.000000	13020.000000	13020.000000	13020.000000
mean	7821.117281	1.406437	3.387219	4.632478	864.000566	66.415054	8.449059	1262.640400	2.573042	3.431235	0.144308	67.293916
std	1884.012138	0.765648	2.087910	2.851356	4044.583167	34.561351	0.499220	730.571224	0.130893	0.805665	0.072319	10.904100
min	4475.000000	0.040500	0.129900	0.113000	0.044500	2.713600	6.647000	41.589200	1.002800	1.208900	-0.012600	41.114400
25%	6191.375000	0.672200	1.758550	2.002425	19.972025	37.579275	8.160600	629.894575	2.494000	2.850375	0.083300	58.393925
50%	7818.750000	1.571650	2.921950	4.539300	66.606950	62.908400	8.269100	1233.866100	2.563500	3.258250	0.154800	69.568250
75%	9449.125000	1.983675	4.789150	6.890750	205.622875	93.816950	8.644925	1891.929100	2.641225	3.956850	0.192100	74.593125
max	11089.500000	4.021900	22.411600	14.421400	40000.000000	258.443600	22.016100	2747.968300	2.997100	9.677000	0.856700	181.575800

Figure 5: Data Description

For the model development, a subset of key features was selected out of the 29 features on the basis of their relevance to well log interpretation. The final input features used include **Depth(ft), Potassium(POTA), Uranium(URAN)(ppm), Gamma Ray(GR)(api), Neutron Porosity(NPHI)(pu), Resistivity(LLD)(ohm), Density(RHOB)(g/cm^3) and Photoelectric Log(PE)(b/e)**.

For training and testing the machine learning models, data from **Block 5, Section 35 called Well A and Block 5, section 16 called Well B** were utilized :

- **Training data** : The dataset from **Well A** provides comprehensive geophysical and operational data. This includes attributes such as **porosity, resistivity , gamma ray** and other well log measurements. These features are pivotal for training models to identify patterns in subsurface formations and predict oil and gas yields.
- **Testing data** : Data from **Well B** is reserved exclusively for testing the models. By segregating training and testing datasets based on geological sections, the study evaluates the model's ability to generalize to new, unseen data from different parts of the basin. This approach ensures robust assessment of the model's performance in predicting sonic logs, reservoir properties and operational efficiencies under distinct geological conditions.

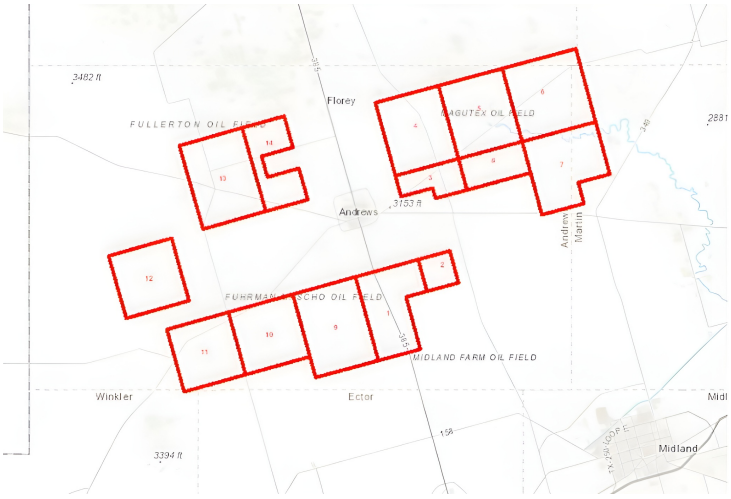

Figure 6: Oil field blocks and sections

The dataset is available through the University Lands Website, which manages mineral rights associated with land in Texas. This platform provides extensive geophysical and geological data from the Midland Basing and other regions across Texas, enhancing accessibility for research and practical applications in the energy sector.

3.2. Tools and Techniques

The different tools and software used in this study utilizes **Python** as the main programming language due to the simplicity, versatility and extensive ecosystem for the machine learning and data analysis. For the machine learning, **Scikit-learn** is used for the traditional models, preprocessing and the evaluation, while **XGBoost** is leveraged for the advanced gradient boosting due to its high efficiency and strong performance at the time of large datasets.

Data manipulation and preprocessing are handled by using the libraries like **Pandas** and **Numpy**, which streamlines the operations on the large and complex datasets. For the visualization purposes, **Matplotlib** and **Seaborn** help in the exploratory data analysis and interpretation of results. And also to manage the noisy and incompleteness of data, **Scikit-learn's** preprocessing tools such as **SimpleImputer** for handling the missing data and feature scaling techniques are also applied to enhance a model's robustness.

The development of the models is carried out in **Jupyter Notebook** for the interactive coding and visualization, while **Visual Studio Code (VS Code)** is used for the debugging part and structured scripting part.

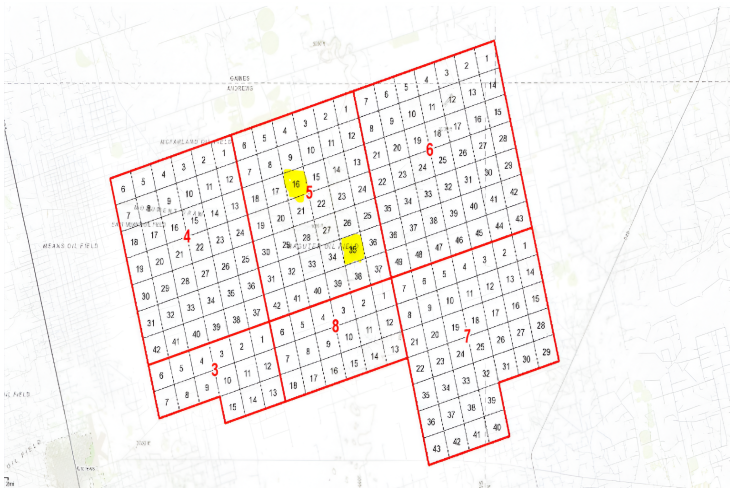

Figure 7: Block 5 and Sections

To handle the well log files in **(.las)** format a python library named **LASIO** is used. This library focus on the loading of the well log data in **(.las)** format and converting it into the **Pandas Data-Frame**, which makes the unstructured data easier to analyze and model development.

In the study, various machine learning models will be employed to predict the sonic logs from the already acquired well logs. The models that are used can be broadly classified into two categories: **Standalone regression models** and **Ensemble models**.

1. **Standalone Regression Models**: Standalone Regression models can be seen as individual predictive models that operate independently without even combining mulitple learning algorithms. The following standalone models are utilized:

    - **Linear Regression (LR)**: It is a fundamental statistical approach that shows the relationship between the dependent and independent variables.
    - **Elastic Net Regression (ENR)**: A regularized regression technique which combines both the Lasso (L1) and Ridge (L2) penalties to improve the generalization ability.
    - **Partial Least Squares Regression (PLSR)**: It is a dimensionality reduction based regression method which finds the fundamental rela-

tions between the predictors and target variables using the latent structures.

- **Support Vector Regressor (SVR)**: A kernel based method which maps input features to a high-dimensional space and finds optimal hyperplane for the regression tasks.
- **K-Nearest Neighbors Regression (KNN-R)** : It is a non parametric algorithm that predict target value on the basis of the average of k-nearest data points in feature space.

2. **Ensemble Models**: Ensemble models can be used to combine multiple decision trees or regression techniques to improve the predictive performance and reduce over-fitting. The following ensemble models were utilized:

3.
- **Random Forest Regressor (RF)**: An ensemble learning model that construct multiple decision trees and make the average of their predictions to improve the accuracy and robustness.
- **XGBoost Regressor**: A gradient boosting framework which optimizes prediction accuracy by iteratively refining the weak learners.
- **Cubist Model**: A rule based model that extends regression trees with the boosting techniques to enhance the accuracy and interpretability.

3.3. Methodology

This methodology involves a systematic approach shown in 8 from data extraction, preprocessing, and handling outliers while both training and testing phases. Methodology enables the evaluation of model robustness and impact of outliers on the predictive performance of models , ultimately helping the development of machine learning models that are resilient and accurate in predicting the logs in oil wells.

1. **Data Extraction**: The study used raw data extracted from the oil wells in the log files. These log files contains the various range of measurements used to determine the subsurface geological properties. Data is extracted from the log files and organized into a structured dataset suitable for the analysis. This step may involve data cleaning , parsing, and preliminary feature selection.

   Python has a robust library named **LASIO** for handling the well log files in **(.las)** format. This library offers the loading of well log data presented

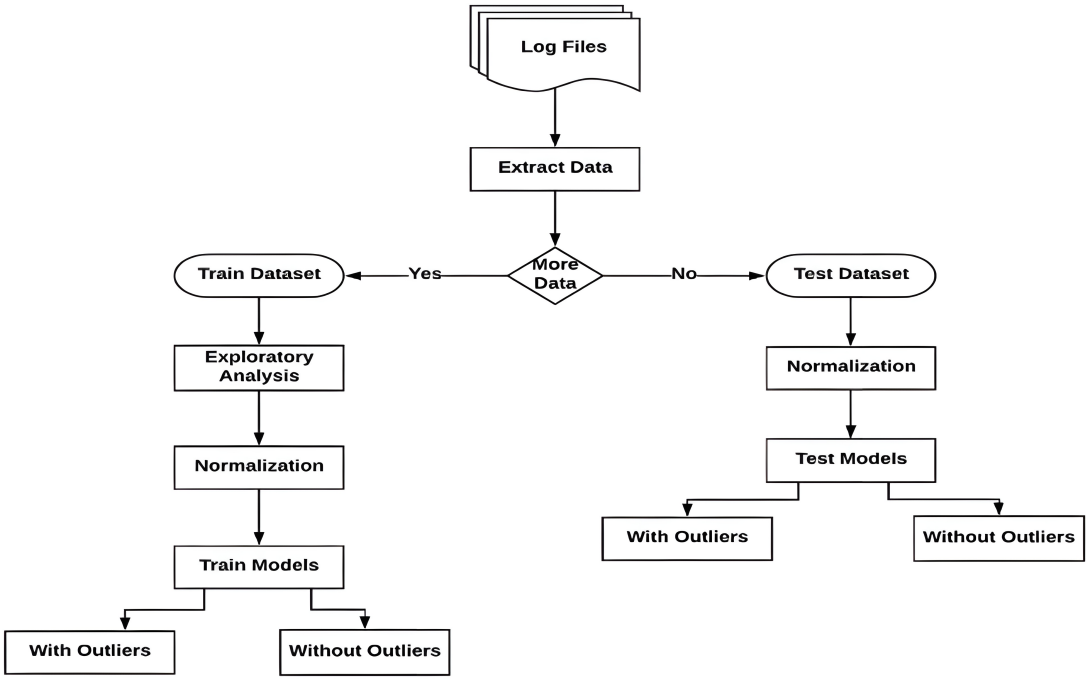

Figure 8: Methodology Flow Chart

in the **.las** format and covert it into a **Pandas Data-frame**, which make the data easier to analyze and use for the development of the models. The same process is used for the test dataset as well.

2. **Exploratory Analysis**: After the extraction part of the data **Exploratory Data Analysis** is crucial for building ML models with oil well log data because it ensure that dataset is clean, the domain specific nature of data and appropriately prepared for the model development. In a domain as complex as petroleum engineering, Exploratory Analysis enables effective selection of features by spotting redundant logs and creating derived features. Results of the model and feature importance can be better interpreted , which can enhance the trust in predictions. Exploratory Analysis helps decide whether to split the data randomly or based on depth of the zones, ensuring the testing and training sets represent the data appropriately.

3. **Data Transformation**: Data Transformation includes converting data in a

suitable format or structure for data analysis and model building process. It ensures the data is compatible with the machine learning algorithms requirements. Data Transformation can be done in many ways like Scaling, Normalization, Log transformation , Power transformation and binning. But during this study first the Skewness of the features is tested. Skewness involves the measurement of the distribution of data around its mean. It helps in identification of as asymmetrical data or wether the data is positively skewed or negatively skewed. The transformation used in the study are following :

(a) **Yeo-Johnson Transformation**: Yeo-Johnson transformation is a technique used to stabilize variance, handle skewness and make data more normal-like, it support both positive and negative values which makes it more versatile.

$$T(y;\lambda) = \begin{cases} \frac{(y+1)^{\lambda}-1}{\lambda}, & \text{if } \lambda \neq 0, y \geq 0 \\ \log(y+1), & \text{if } \lambda = 0, y \geq 0 \\ \frac{-(|y|+1)^{2-\lambda}+1}{2-\lambda}, & \text{if } \lambda \neq 2, y < 0 \\ -\log(|y|+1), & \text{if } \lambda = 2, y < 0 \end{cases}$$

(b) **Box-Cox Transformation**: Box-Cox transformation is also used to stabilize variance and reduce skewness. It is mainly useful when the data contain positive values and is heavily skewed.

$$T(y;\lambda) = \begin{cases} \frac{y^{\lambda}-1}{\lambda}, & \text{if } \lambda \neq 0 \\ \log(y), & \text{if } \lambda = 0 \end{cases}$$

(c) **Min-Max Scaling**: Min-Max scaling is a normalization technique used to rescale the feature into a specific range, range is usually between 0 and 1. It is commonly used when feature have various units or scales.

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

4. **Handling Outliers and Data Splitting**: Outliers in dataset are identified on the basis of the difference between the observed caliper measurement in the logs and 7.88 reference value, which is predefined calibration measure of caliper tool. Observation which shows a deviation from 7.88 are considered as outliers and are removed after. Data Splitting is done on the basis of how the outliers are handled which result in the creation three distinct datasets:

(a) **x-all and y-all**: This dataset involves all the original data, outliers are also included in these datasets after the splitting.

(b) **x-small and y-small**: This dataset includes making the resultant dataset by removing only the observations where the caliper measurements were taken as outliers.

(c) **x-no-out and y-no-out**: The dataset is created by removing the outliers from all the other logs, not just from the caliper measurements.

5. **Model Training and Testing**: After the splitting of the dataset into three subsets (original data, data after removing caliper outliers, and data with no outliers), the model training and testing process starts. Each model is then trained and validated on all the three subsets to evaluate their robustness and adaptability.

   During the training period, models undergo the hyperparameter tuning part and are assessed using the performance metrics like **RMSE, MAE and** R^2. To ensure the reliable model performance, **k-fold cross-validation** is applied, where the data is divided into k subsets and model is trained and validated k times, each time a different subset is used for validation while the remaining are used for training which helps in reducing the over-fitting problem.

   The testing phase also closely mirrors the training process, ensuring the consistency through normalization. Testing is done under two conditions:

   (a) **Testing with Outliers**: This scenario assesses the model performance in the real-world conditions, where datasets may contain some anomalies or some noisy values. It provides the insights into how well models generalize when are faced with the imperfect data.

   (b) **Testing without Outliers**: By evaluating the models on a clean dataset, this approach determines sensitivity of the models to outliers and whether their accuracy while predicting improves in the absence of anomalies.

   The evaluation of the models is done by the following metrices:

   (a) **Root Mean Square Error (RMSE)**:

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

(b) **Mean Squared Error (MSE)**:

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

(c) **R-squared Error (R^2)**:

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

4. Experiments

The study used raw data extracted from the oil wells in the log files. These log files contains the various range of measurements used to determine the subsurface geological properties. Data is extracted from the log files and organized into a structured dataset suitable for the analysis. This step may involve data cleaning , parsing, and preliminary feature selection.

Python has a robust library named **LASIO** for handling the well log files in **(.las)** format. This library offers the loading of well log data presented in the **.las** format and covert it into a **Pandas Data-frame**, which make the data easier to analyze and use for the development of the models. The same process is used for the test dataset as well. The summary of the dataset is shown in the 9.

	DEPT	POTA	URAN	THOR	LLD	GR	CALI	BHVT	RHOB	PE	NPHI	MDT
count	13020.000000	13020.000000	13020.000000	13020.000000	13020.000000	13020.000000	13020.000000	13020.000000	13020.000000	13020.000000	13020.000000	13020.000000
mean	7821.117281	1.406437	3.387219	4.632478	864.000566	66.415054	8.449059	1262.640400	2.573042	3.431235	0.144308	67.293916
std	1884.012138	0.765648	2.087910	2.851356	4044.583167	34.561351	0.499220	730.571224	0.130893	0.805665	0.072319	10.904100
min	4475.000000	0.040500	0.129900	0.113000	0.044500	2.713600	6.647000	41.589200	1.002800	1.208900	-0.012600	41.114400
25%	6191.375000	0.672200	1.758550	2.002425	19.972025	37.579275	8.160600	629.894575	2.494000	2.850375	0.083300	58.393925
50%	7818.750000	1.571650	2.921950	4.539300	66.606950	62.908400	8.269100	1233.866100	2.563500	3.258250	0.154800	69.568250
75%	9449.125000	1.983675	4.789150	6.890750	205.622875	93.816950	8.644925	1891.929100	2.641225	3.956850	0.192100	74.593125
max	11089.500000	4.021900	22.411600	14.421400	40000.000000	258.443600	22.016100	2747.968300	2.997100	9.677000	0.856700	181.575800

Figure 9: Data description

The well of Block 5, Section 35 used for the training purpose. As based on the already existing guidelines in petroleum engineering log analysis, the logs used in this study will be:

- **CALI** : It measures the borehole caliper readings, but this log is dropped from the study after ensuring their is no chance of washout .
- **TENS** : TENS represents the tension logs and is not required further in the study.

- **DRHO, XPHI, SPHI, DPHI, NDSN, FDSN, ITTT** : These logs are calculated from the recorded logs and hence are removed to avoid redundancy.
- **MSFL and LLS** : These are the resistivity logs used to measure resistivity, LLD (Deep Resistivity) will be retained as this log provide data from the uninvaded zones, which makes the log more reliable for analysis.
- **AHVT** : Logs like AHVT indicates the volume of the well for cement placement and it is unrelated to the prediction variable.
- **GRTO, GRTH, GRKT, GKUT** : Are the Gamma ray logs which measures the various gamma measurements. GKUT and GRTO measure only the same parameter, but GKUT is mostly unavailable in many logs, so these logs will be excluded.
- **MDT (Mono Delta T** : The target variable of the analysis is MDT. This log will be used as the dependent variable in model building.

First, to assess the missing values in the well log data, a function was created to identify the gaps in the log readings. This function basically scans the data to detect any missing values across different logs, providing some insights into data completeness and highlighting specific ranges of depth where data is absent. Upon applying this function, it is observed that missing values were primarily concentrated in the well log, mainly in the 'NPHI', 'RHOB', 'GR' , 'CALI' and 'MDT' logs. These gaps shows discontinuities in the log measurements, which can likely caused by sensor malfunctions, well conditions or incomplete data acquisition.

To further investigate the gaps in the data, a **visual inspection** of log data was conducted as shown in 10. This confirmed that missing values were localized in the specific depth depth intervals, by reinforcing need for the appropriate handling strategies. After the careful inspection, we decided to **drop the affected data points** by maintaining the integrity of the data. This decision was done on the basis of the relatively small percentage of missing values and potential impact on the model's performance.

Secondly, it is checked whether are their any chances of washout. Washout occurs when the borehole is enlarged beyond the original size drilled be the tool. Main factors contributing in the washout are rock formations or high drilling fluid pressure. This can further lead to inaccurate caliper too readings. To detect washout caliper readings are compared to the borehole size. A washout can be suspected if : **Caliper reading > 1.5 * Set borehole size**

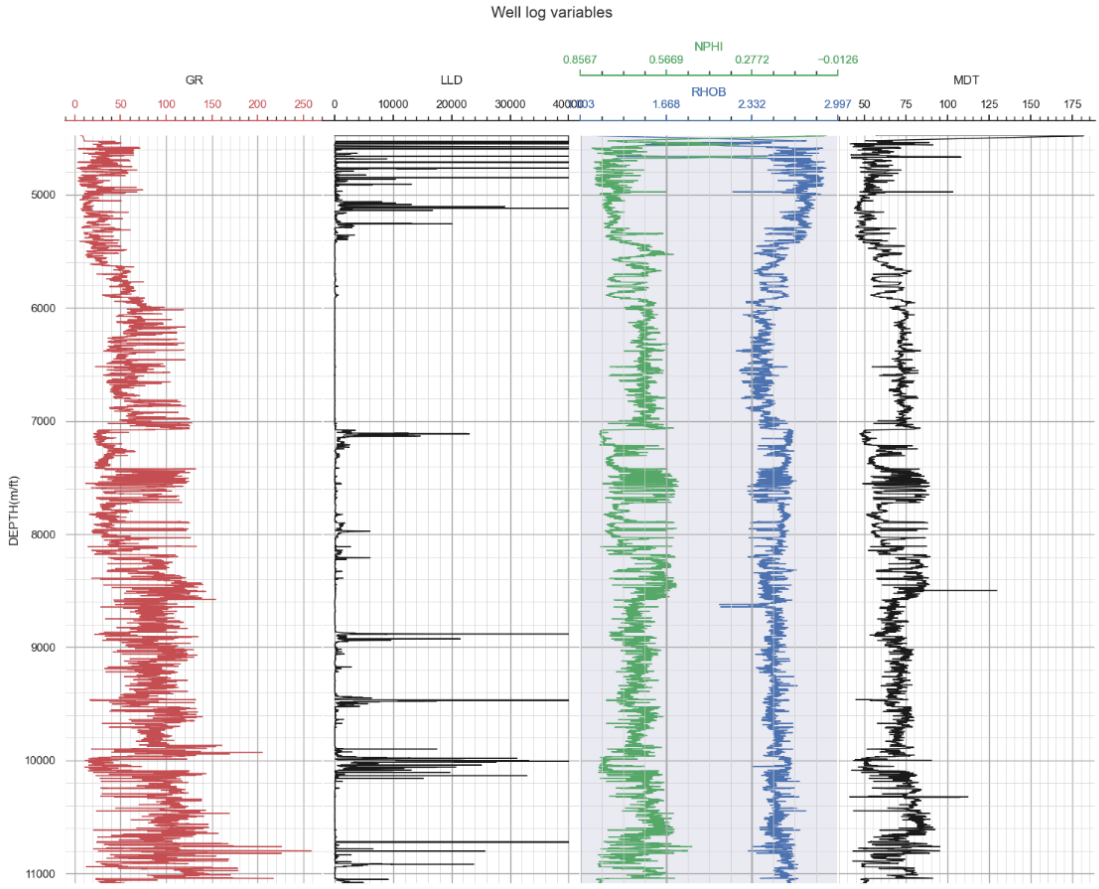

Figure 10: well log variables

From the data collected from the Section 35 and the wire-line plot of CALI log above it can be seen that the maximum caliper readings of 22.4116 exceeds the threshold value of 11.82 , which confirms the presence of washout as also shown in the wire-line plot of the **CALI** variable as shown in 11.

To see the statistical view of the whole data a heat-map is generated which shows how much a log variable can effect the other variable or how one variable is correlated to another variable as shown in 12. From the 12 it is clear that Well Depth and Volume variables do not exhibit a clear linear relationship with the sonic log. These logs are also highly collinear, which makes them redundant predictors. In conclusion, Volume will be dropped, as it logically varies with depth variable.

If we look at the the distribution of other logs with respect to the MDT log

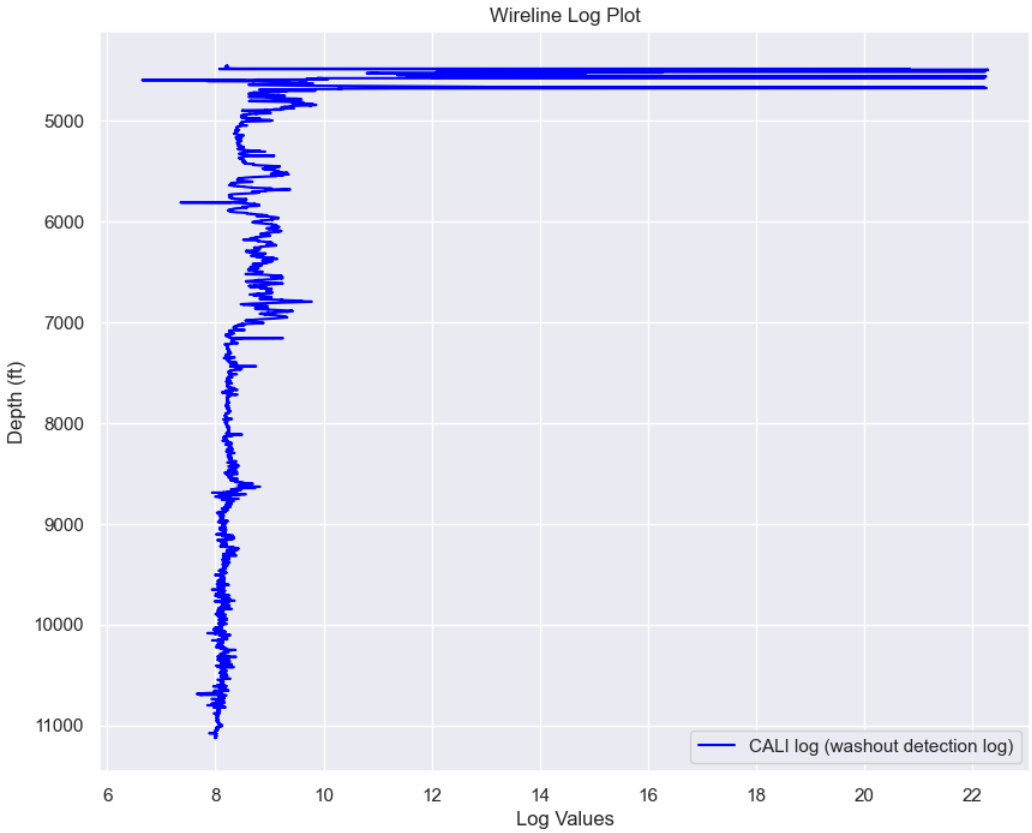

Figure 11: CALI wire-line plot

variable in the 13 . As we shown that the uni-variate distribution of resistivity is heavily skewed. The relationship of resistivity with sonic log appears non-linear, necessitating further transformations. Postassim, Uranium, Thorium, Gamma Ray and Neutron Porosity show a positive correlation with Sonic log while Density and Photo Electric Logs exhibit a negative correlation with Sonic log.

Resistivity, Uranium and Gamma Ray showed skewed distribution and needs to undergo transformation. First the skewness will be tested and all the skewed features will be transformed using the **Box-cox Transformation** to ensure no feature has more influence than other due to its values. Neutron Porosity contains negative values, so **Yeo-Johnson transformation** is used, as it can handle both positive and negative values. **Depth** exhibits values that are significantly larger as compared to other variables. To ensure the consistent scaling and to prevent the Depth from dominating model learning, a **Min-Max scaler** is used, which gives

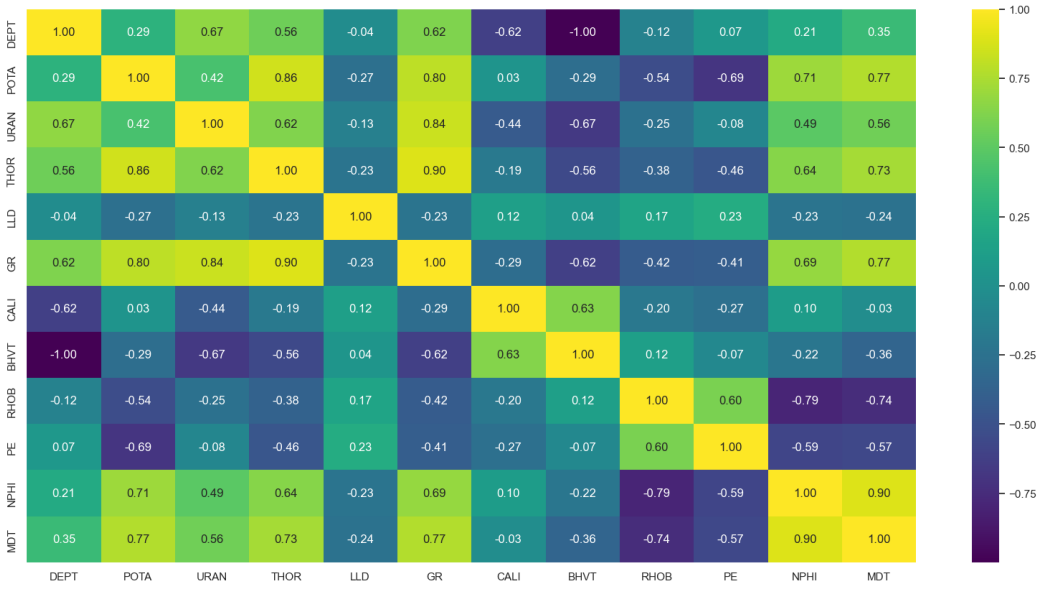

Figure 12: Heat-map of log attributes

the values ranging between 0 and 1. Potassium does not exhibit any noticeable skewness and values are within a reasonable range, hence no transformation is necessary to apply.

From the above figure it can be seen that the dataset exhibits a mixture of lithologies with transitions between shale, sandstone and limestone, as inferred from gamma ray, resistivity and porosity indicators. The low neutron porosity and high resistivity zones point to dense, compact formations, whereas zones with high neutron porosity and gamma ray values suggest shale or less compact formations.

To address the potential outliers in the data, observations were removed based on the difference between recorded caliper measurements in the logs and the standard value 7.88. Three dataset are prepared based on this which is already discussed in the methodology section.

- **x-all and y-all**
- **x-small and y-small**
- **x-no-out and y-no-out**

The outliers presented in the different logs can be seen using the visualization methods as shown 15 or by using the IQR Range method where Inter-quartile

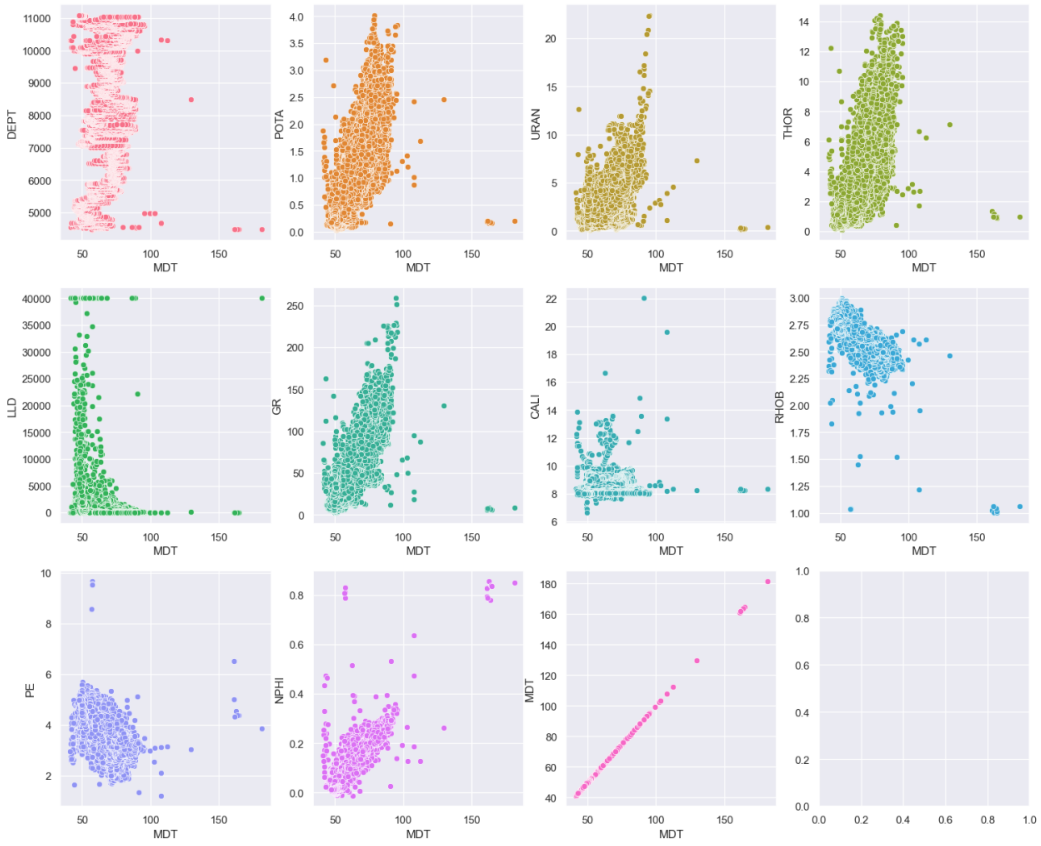

Figure 13: Logs Distribution

range is calculated of each numerical column in given dataset. IQR is a measure of statistical dispersion which computes the lower bound and upper bound for outlier detection.

The lower bound and upper bounds for the logs presented in dataset are following :

- DEPT : [1304.75, 14335.75]
- POTA : [-1.29501250, 3.95088750]
- URAN : [-2.78735000, 9.33505]
- THOR : [-5.3300625, 14.2232375]
- LLD : [-258.5042499996, 484.009915]

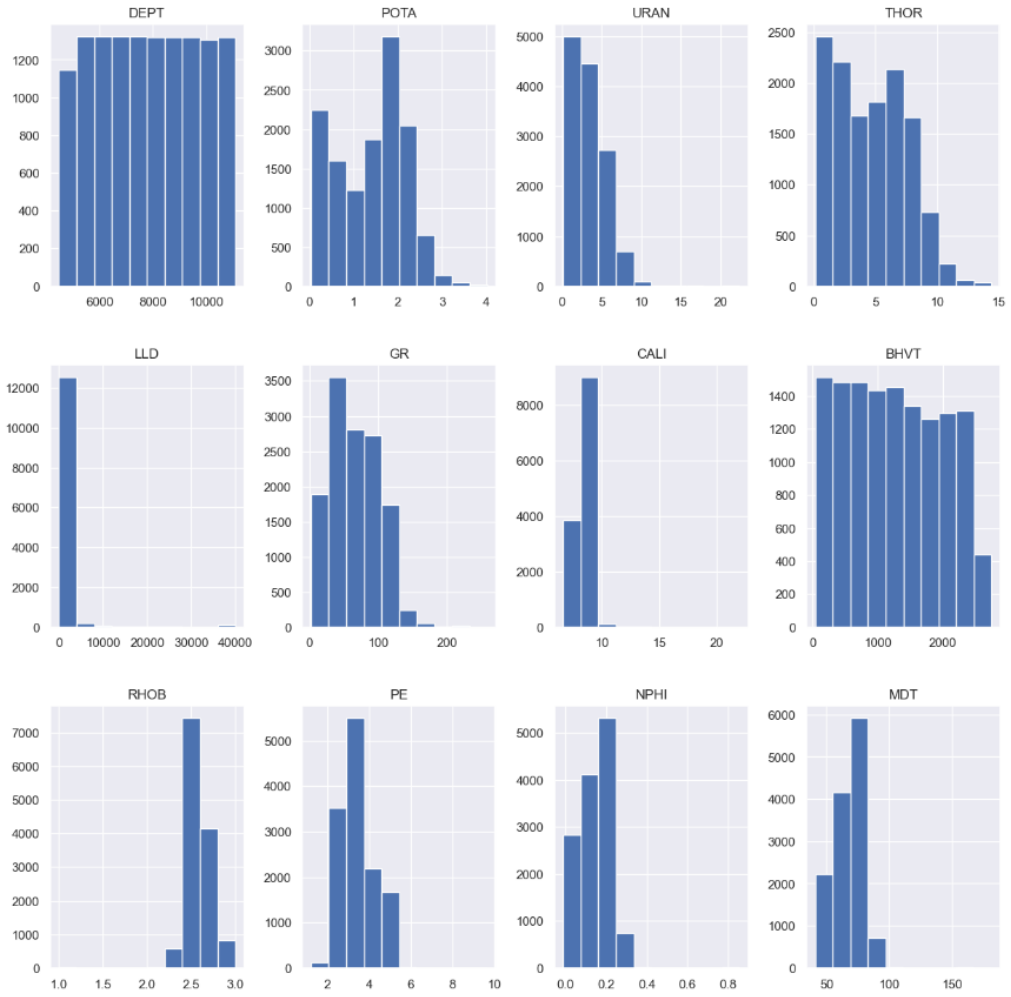

Figure 14: logs parameters distribution

- GR : [-46.7772375001, 178.173462500003]
- CALI : [7.4341125, 9.371212500002]
- RHOB : [2.27316249999998, 2.86206250000004]
- PE : [1.19066249999998, 5.61656250000001]
- NPHI : [-0.07989999999999998, 0.3552999999995]
- MDT : [34.0951249999999, 98.891925000001]

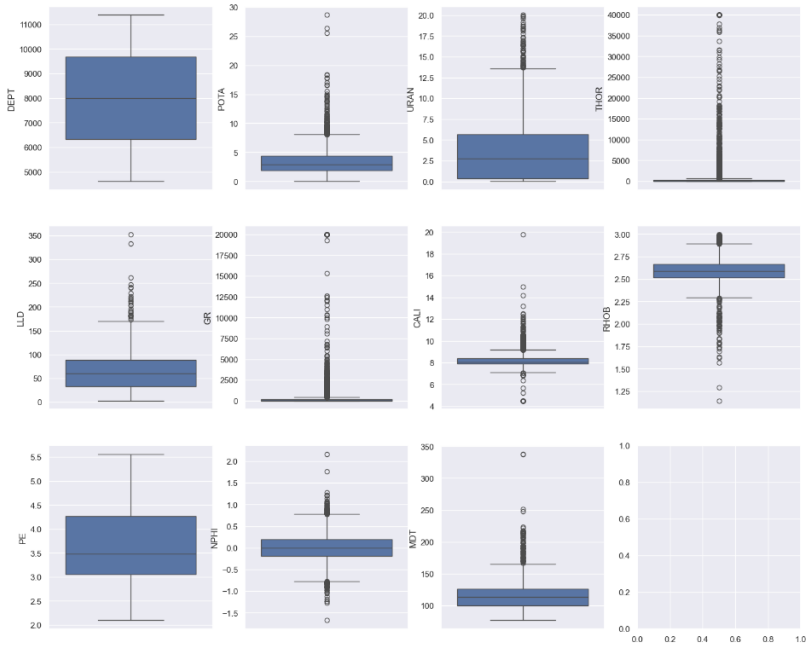

Figure 15: Visual Inspection of outliers of different logs using box-plot

Now as it is know that outliers will be dropped which are not in the IQR range of the log. Due to dropping the outliers from the dataset the shape of data is changed. Shape of data-frame with caliper outliers removed is [12673, 11] and Shape of data-frame will all log outliers removed is [10360, 11]. The effect of the outlier removal can be easily shown on the box-plots created after shown in 16. Now we have to see the effect of the outlier removal on the model's performance and accuracy by applying the standalone and ensemble models on the prepared data.

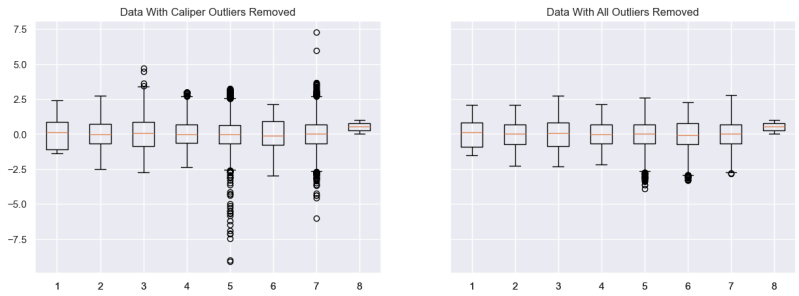

Figure 16: Before and After of outlier removal

5. Results and Discussion

In this study, a very diverse range of machine learning models were utilized to predict the sonic logs form pre-existing well log data collected in Midland Basin. As already discussed the training data was sourced from Block 5, Section 35, while testing was performed using data from the Block 5, section 16, which ensures the spatial and geological variability between the datasets. To ensure a robust evaluation, the model were validated using **k-fold Cross-validation**. Which involves partitioning the training dataset into k subsets (fold) of equal size. The model is trained on the k-1 folds and validated using the remaining folds, rotating the validation folds across all subsets.

The Comparison of the models was based on metrics such as Mean absolute error (MAE) , Root Mean Square Error (RMSE) and R-squared. The overall results shows that the ensemble models , specially Random Forest, XGBoost and Cubist performed better than the simpler approaches in generalizing to the testing dataset, which demonstrate their utility in sonic log prediction and also of reservoir characterization.The results that are generated over data is divided into the three parts where first part represent the **Performance of models on original data** . The results table for the different models performance on the original data is given in 2.

After removing the caliper outliers the model performed differently and their was slightly increase in the performance of the models. Models are able to catch the data more efficiently after removing the caliper outliers from the data. The results table for the model performance on the data after removing the caliper outliers is given in 3.

In the third table all the outliers presented in the log data are fixed and now after getting the cleanest form of the data. All the models are again used to see whether

Machine Learning Models	Mean Absolute Error (MAE)	Root Mean Squared Error (RMSE)	R-Squared (R2)
Linear Regression	2.42	3.9191	0.8710
Elastic Net Regression	2.39	3.8858	0.8736
Partial Least Squares	2.39	3.8858	0.8710
KNN Regressor	0.98	2.1863	0.9593
Support Vector Regressor	1.21	2.1000	0.9617
Random Forest	0.45	1.0099	0.9913
Cubist Model	1.02	2.2271	0.9581
XGBoost Regressor	1.02	1.4798	0.9810

Table 2: Models Performance on all Original Data

their is any change on the performance of the models or their is any improvement in recognizing the patterns of data and to see predictive behavior of the logs. The results table for no outlier data is given in table4. In table 4 the final performance evaluation of the all the models applied on the data is provided.

To discuss the result it can be seen that the **Linear Regression** servers here as a baseline model, yielding an MAE of 2.01 and RMSE of 3.1464. Its R2 value 0.9060 indicates that while model explain 90.6 percentage of the variance in the data, its predictive accuracy is limited due to relatively high error metrics.

The other model **Elastic Net Regression** slightly improves upon Linear Regression with an MAE of 1.98 and reduced RMSE value of 2.7676. The R2 value of 0.9079 suggests a marginally better fit to the data, showcasing Elastic Net's ability to handle feature selection and multicollinearity effectively.

The third model **Partial Least Squares** comparably to Elastic Net, with an MAE of 1.99, an RMSE of 2.7676 and an R2 of 0.9062. This similarity between

Machine Learning Models	Mean Absolute Error (MAE)	Root Mean Squared Error (RMSE)	R-Squared (R2)
Linear Regression	2.27	3.5986	0.8906
Elastic Net Regression	2.25	3.5707	0.8973
Partial Least Squares	2.25	3.5707	0.8919
KNN Regressor	0.93	1.8547	0.9703
Support Vector Regressor	1.18	2.034	0.9639
Random Forest	0.44	0.9380	0.9924
Cubist Model	0.94	1.8574	0.9698
XGBoost Regressor	0.98	1.4142	0.9826

Table 3: Models Performance on Data after removing caliper outlier

Machine Learning Models	Mean Absolute Error (MAE)	Root Mean Squared Error (RMSE)	R-Squared (R2)
Linear Regression	2.01	3.1464	0.9060
Elastic Net Regression	1.98	2.7676	0.9079
Partial Least Squares	1.99	2.7676	0.9062
KNN Regressor	0.92	1.5874	0.9687
Support Vector Regressor	1.11	1.7549	0.9619
Random Forest	0.42	0.7	0.9938
Cubist Model	0.90	1.4456	0.9742
XGBoost Regressor	0.89	1.2569	0.9804

Table 4: Models Performance after removing all outliers from the Data

both the models highlights its utility in reducing dimensionality for the datasets with predictors.

KNN model demonstrate a marked improvement over linear models, with an MAE of 0.92 and an RMSE of 1.5874. R2 value of 0.9687 indicates a strong predictive ability. The model's success likely stems from its flexibility in capturing local data patterns, although the model's performance can depend heavily on the choice of hyper-parameters like number of neighbors.

Next model **Support Vector Regressor** achieves an MAE of 1.11 and an RMSE of 1.7549 , with an R2 value of 0.9619. While its error metrics are sightly higher than those of KNN, SVR still delivers more robust predictions by finding a between the bias and variance through its kernel-based approach.

Now if we look at the ensemble model as compare to the standalone models **Random Forest** outperforms all the other models, by achieving the lowest error metrics (MAE = 0.42 , RMSE = 0.7) and the highest R2 value of 0.9938. This indicates the exceptional predictive accuracy and suggests that Random Forest effectively shows complex patterns in the data through ensemble learning and feature averaging.

The dotted line in the below figures represent the ideal predictions of data. while 'blue' dots represent the values predicted by using original data before transformation and scaling, 'yellow' dots shows the predicted value of data after removal of caliper outliers and the 'pink' dots represents predicted values with no outliers as shown in 17.

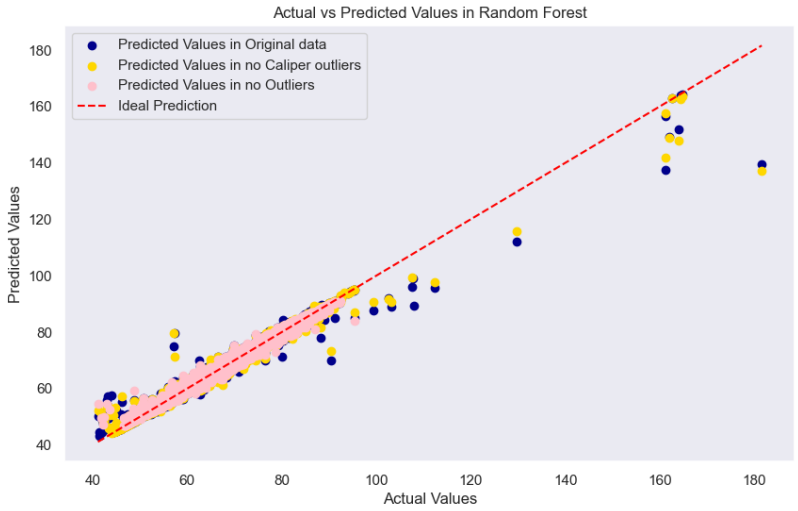

Figure 17: Actual vs Predicted Cross-plot Random Forest

The **Cubist Model** also performs very well, with an MAE of 0.90, an RMSE of 1.4456 and R2 value of 0.9742. Its good performance demonstrate the model's strength in handling both the linear and non-linear relationships, making it a competitive alternative to Random Forest.

The 18 below represent the distribution of predicted data points using Cubist model where 'red' dots are data points of original data, 'violet' dots belongs to data point with no caliper outliers and 'blue' represents data points having no outliers.

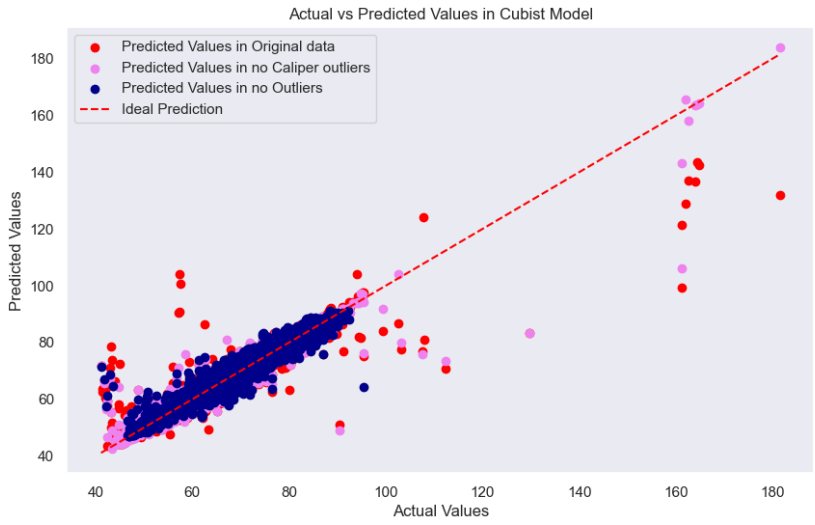

Figure 18: Actual vs Predicted Cross-plot Cubist

The Final model **XGBoost Regressor** also delivers impressive results, with an MAE of 0.89, RMSE of 1.2569 and R2 of 0.9804. As being a gradient-boosting algorithm, XGBoost Regressor effectively minimizes error through sequential learning , placing it between the top-performing models in this study.

For a better understanding of impact of using XGBoost Regressor on dataset a visual graph of predicted data points is crafted as show in19 where each different colors represents different predicted data points like 'light blue' points shows original data points , 'green' one shows points with no caliper outliers and 'purple' points belongs to the data with no outliers (cleaned data).

If we summarize the results, the ensemble models demonstrated superior performance, with Random Forest achieving the very best results (MAE = 0.42, RMSE = 0.7, R^2 = 0.9938), showcasing its ability to capture complex data patterns effectively as shown in 20. The cubist model also performed well (MAE = 0.90, RMSE = 1.4456, R^2 = 0.9742), proving its robustness in handling both linear and non-

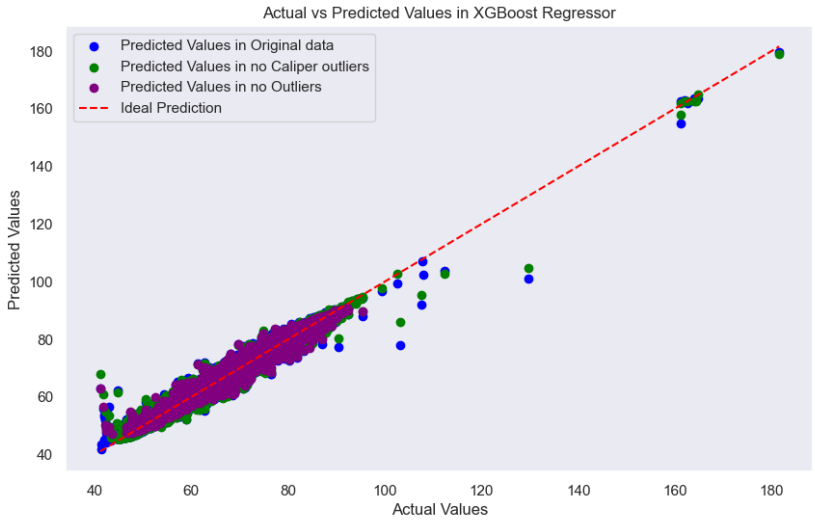

Figure 19: Actual vs Predicted Cross-plot XGBoost Regressor

linear relationships. Similarly, XGBoost Regressor delivered strong results (MAE = 089, RMSE = 1.2569, R^2 = 0.9804), leveraging gradient boosting to efficiently minimize errors.

Also by looking at all the cross plots of Random Forest, Cubist and XGBoost it can be seen that while using original data the predicted values were a bit scattered across the graph in all the models, after the removal of caliper outliers the predicted data points can be seen closer to ideal prediction line which is a very good sign and when we use cleaned data (data with no outliers) predicted values are no longer scattered and can be seen sticking to the ideal line which make the residual error very low and increase the model's performance and efficiency.

The sonic or acoustic log measures the time it takes for an elastic wave to travel through a formation. The speed of wave propagation through the formation is influenced by factors such as lithology , rock texture and porosity. The wave velocity tends to decrease with higher porosity and increase with lower porosity or high rock density. High velocity zones may correspond to dense or compact regions, while low-velocity zones may indicate areas with higher porosity. The graph below represents logs for resistivity, gamma ray, MDT (sonic) , and neutron porosity across various depths in a well , likely to identify the characteristics of lithology, porosity and subsurface.

The analysis of the above well log graphs, including resistivity, gamma ray , MDT (sonic) and neutron porosity, provides valuable insights into subsurface

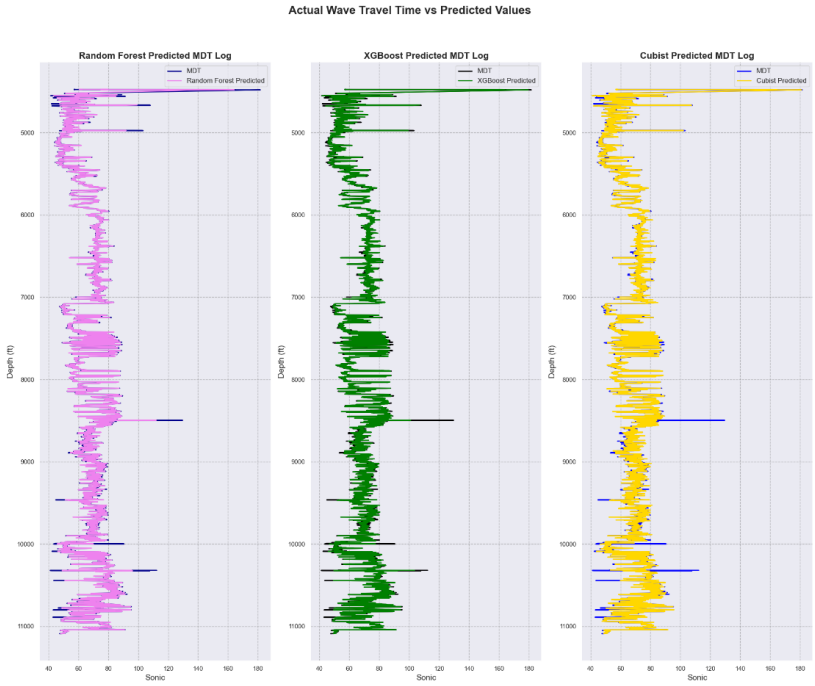

Figure 20: Actual wave vs Predicted values using ensemble models

porosity, lithology and composition across various depths From 5000 ft to 6000 ft, formation is likely dominated by porous sandstones. This interpretation is supported by low resistivity values (high initially but decrease as reaching near 5000 ft) , which indicates less dense material and low gamma ray readings (0-50 API) , reflecting minimal radioactivity. Additionally , high neutron porosity in this zone suggests significant porosity, while higher sonic travel time shows lower wave velocity, further confirming the presence of a porous sandstone formation.

Between 6000 ft and 8000 ft , the lithology transitions to a mix of sandstone, limestone and shale. Moderate resistivity values point to denser materials, while increasing gamma ray readings indicate the presence of shale interbedded with sandstone and limestone. In this interval, decreasing neutron porosity reflects reduced porosity due to compaction and decreasing sonic travel time (showing higher velocity) aligns with the presence of denser lithologies like limestone or compact sandstone.

From 8000 ft to 11000 ft, the formation is likely dominated by shale or sandy-shale layers with varying porosity. Fluctuations in resistivity within this interval suggest changes in compaction and porosity. Higher gamma ray readings indicate

a shale-rich composition, while slight increases in neutron porosity correspond to less compact zones. Sonic travel time also increases at certain depths, indicating lower wave velocities consistent with shale or sandy-shale formations.

The difference between the impact of different standalone models and ensemble models predictions can be easily seen in the actual vs predicted cross-plot 21. As it can be seen the standalone model's prediction is more scattered across while the ensemble model predicted values is sticking along with the ideal or perfect prediction line which shows us how well ensemble models are capable of the reading various patterns of the data and predicting the values based on that analysis and pattern evaluation. The residual error like Mean Absolute Error or Root Mean Square Error is also very low for ensemble group as compared to the standalone regression models.

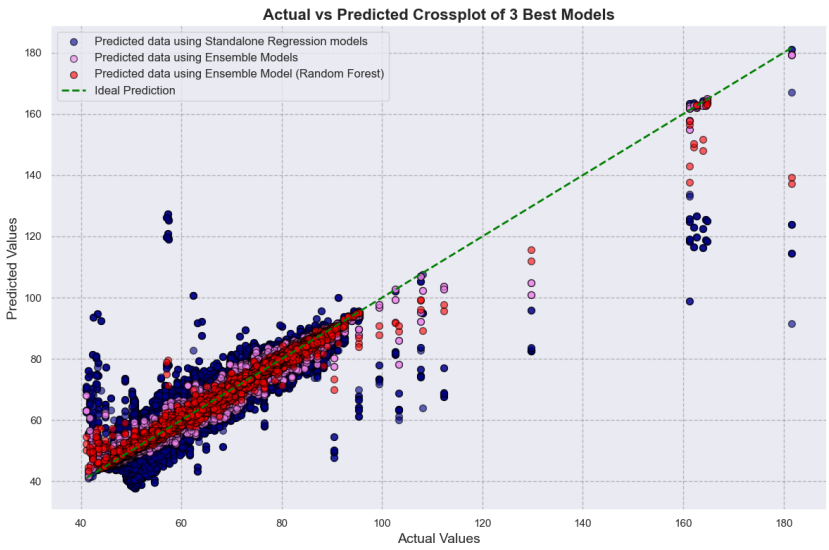

Figure 21: Standalone vs Ensemble Prediction Cross-plot

6. Conclusion

After the brief experiments and analysis this study successfully demonstrates the potential of machine learning models in the prediction of sonic logs using existing well log data gathered from the Midland Basin (Texas, USA). The research aim towards evaluating how well data collected from one section of a geological block can be used to predict logs in another , emphasizing the importance of spatial and geological variability. The framework developed identifies the ideal combination of input logs, revealing that **Density (RHOB), Resistivity (LLD) and Neutron**

porosity (NPHI) are the most critical variables for accurate prediction. These logs are very strongly associated with lithology, rock texture and porosity, which play an important in influencing sonic travel time. Ensemble models, including Random Forest , XGBoost and Cubist emerged as top performers in this study, with Random Forest having the best results (MAE = 0.42 , RMSE = 0.7 , R^2 = 0.9938). These models demonstrated exceptional capability in capturing complex patterns within the data as seen in the figure below, outperforming simpler models such as linear regression and Support vector regressor.

Preprocessing steps such as removal of the caliper outliers and extensive data cleaning , further successfully enhanced the model performance by reducing residual errors and aligning predicted values closer to ideal predicted line. The study also showcased the ability of machine learning models to generalize effectively across datasets with different geological characteristics, reinforcing their robustness and adaptability for different sections within the basin.

In addition to the quantitative performance, the models provided valuable geological insights. For instance, analysis of well log data across depths (5000 to 11,000 ft) can be seen in graph below highlights lithological transitions from porous sandstone to compact limestone and layers of shale, with sonic travel time variations aligning well with these geological features. High neutron porosity and low resistivity were linked to the formation of porous, while increasing gamma-ray readings and lower porosity levels corresponded to denser and compact lithologies. These observations validates the ability of machine learning model to incorporate geological complexities into their predictions, further strengthening their utility in subsurface characterization.

This study has significant cost-saving implications for the oil and gas industry. Acquiring new well log measurements, specifically sonic logs can be expensive and time-intensive, which requires specialized tools and equipment. By leveraging existing well log data, the proposed machine learning framework reduce the need for additional field measurements, making it a cost effective alternative for reservoir characterization. Predicting Sonic logs from already available data can also optimize resource allocation and streamline exploration and production activities. The lower dependence on direct measurements enables operators to make timely and informed decisions, which leads to better economic efficiency in hydrocarbon exploration.

Furthermore , the insights generated by this study are highly beneficial for understanding and managing hydrocarbon reservoirs. Accurate prediction of sonic logs helps in identifying subsurface properties such as porosity, lithology and fluid content, which are critical for reservoir characterization. The ability of models to

delineate compact formations, high-porosity zones and transitions between different lithologies provides valuable information for estimation of reservoir quality and hydrocarbon potential. This predictive capability support enhanced oil recovery (EOR) strategies, reservoir modeling and efficient field development planning. Future research could extend this approach by incorporating advanced deep learning techniques, integrating additional datasets and apply these models to other geological settings, further enhancing the utility of machine learning in addressing the complex challenges in subsurface exploration, prediction of DTC sonic logs or DTS sonic logs and hydrocarbon reservoir management.

References

1. Agung, L., Sastranagara, T., Fitriah, A., Hastuti, P. & Raharjo, I. Sonic Log Derived Porosity In Kamojang Geothermal Field (Aug. 2016).
2. Hua, W. & Yushun, Z. Research status and prospect of artificial intelligence in logging data processing and interpretation. *Well Logging Technol* **45,** 345e56 (2021).
3. Li, H. & Misra, S. Long short-term memory and variational autoencoder with convolutional neural networks for generating NMR T2 distributions. *IEEE Geoscience and Remote Sensing Letters* **16,** 192–195 (2018).
4. Chopra, S., Sharma, R. K., Marfurt, K. J., Zhang, R. & Wen, R. Improving porosity and gamma-ray prediction for the Middle Jurassic Hugin sandstones in the southern Norwegian North Sea with the application of deep neural networks. *Interpretation* **10,** T25–T34 (2022).
5. Meshalkin, Y., Shakirov, A., Popov, E., Koroteev, D. & Gurbatova, I. Robust well-log based determination of rock thermal conductivity through machine learning. *Geophysical Journal International* **222,** 978–988 (2020).
6. Bukar, I., Adamu, M. & Hassan, U. *A machine learning approach to shear sonic log prediction* in *SPE Nigeria Annual International Conference and Exhibition* (2019), D023S026R001.
7. Cranganu, C. & Bautu, E. Using Gene Expression Programming to estimate sonic log distributions based on the natural gamma ray and deep resistivity logs: A case study from the Anadarko Basin, Oklahoma. *Journal of Petroleum Science and Engineering* **70,** 243–255 (2010).

8. Cranganu, C. & Breaban, M. Using support vector regression to estimate sonic log distributions: A case study from the Anadarko Basin, Oklahoma. *Journal of Petroleum Science and Engineering* **103,** 1–13. ISSN: 0920-4105. https://www.sciencedirect.com/science/article/pii/S0920410513000430 (2013).

9. Nyein, C. Y., Hamada, G. M. & Elsakka, A. *'Artificial neural network (ANN) prediction of porosity and water saturation of shaly sandstone reservoirs* in *Proc. AAPG Asia Pacific Region, 4th AAPG/EAGE/MGS Myanmar Oil Gas Conf., Myanmar, Global Oil Gas Hotspot, Unleashing Petroleum Syst. Potential* (2018).

10. Cao, J., Shi, Y., Wang, D. & Zhang, X. Acoustic Log Prediction on the Basis of Kernel Extreme Learning Machine for Wells in GJH Survey, Erdos Basin. *Journal of Electrical and Computer Engineering* **2017,** 3824086. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1155/2017/3824086. https://onlinelibrary.wiley.com/doi/abs/10.1155/2017/3824086 (2017).

11. Tariq, Z., Elkatatny, S., Mahmoud, M. & Abdulraheem, A. *A new artificial intelligence based empirical correlation to predict sonic travel time* in *International Petroleum Technology Conference* (2016), D012S057R001.

12. Pandey, S. & Saraiya, R. *Prediction of Sonic Log Data Using Machine Learning Regression Methods* in *Marine Acquisition Workshop 2018* (2018), cp–560.

13. Anifowose, F. A., Labadin, J. & Abdulraheem, A. Ensemble machine learning: An untapped modeling paradigm for petroleum reservoir characterization. *Journal of Petroleum Science and Engineering* **151,** 480–487. ISSN: 0920-4105. https://www.sciencedirect.com/science/article/pii/S0920410517300712 (2017).

14. Sun, J. & Li, H. Financial distress prediction using support vector machines: Ensemble vs. individual. *Applied Soft Computing* **12,** 2254–2265. ISSN: 1568-4946. https://www.sciencedirect.com/science/article/pii/S1568494612001263 (2012).