Machine Learning Approach for Vertical Permeability Prediction and Permeability Anisotropy of Hugin Sandstone Formation (Volve field)

Thesis submitted in partial fulfilment

for the Award of

Master of Technology

in

Petroleum Engineering

by

VIKRAM KUMAR

Department of Petroleum Engineering and Geoengineering

RAJIV GANDHI INSTITUTE OF PETROLEUM TECHNOLOGY

JAIS, INDIA – 229304

"MPE19-005" **"2021"**

THESIS APPROVAL SHEET

"***Machine Learning Approach for Vertical Permeability Prediction and Permeability Anisotropy of Hugin Sandstone Formation (Volve field)"***

By

Vikram Kumar

(MPE19-005)

A Thesis Approved

By

THESIS COMMITTEE

Dr. Satish Kumar Sinha
(Thesis Supervisor)
Associate Professor,
Dept. of Petroleum Engineering and
Geoengineering

Dr. Alok Kumar Singh
(Co-supervisor)
Associate Professor,
Dept. of Petroleum Engineering and
Geoengineering

Dr. Shivanjali Sharma
(DPGC)
Assistant Professor,
Dept. of Petroleum Engineering and
Geoengineering

Dr. Amit Kumar
(Member)
Assistant Professor,
Dept. of Petroleum Engineering and
Geoengineering

Dr. Amit Saxena
(Member)
Assistant Professor,
Dept. of Petroleum Engineering and
Geoengineering

Dr. Manoj Kumar Rajpoot
(External Member)
Assistant Professor,
Dept. of Mathematical Sciences

CERTIFICATE

It is certified that the work contained in the thesis titled ***Machine Learning Approach for Vertical Permeability Prediction and Permeability Anisotropy of Hugin Sandstone Formation (Volve field)*** *by* ***Vikram Kumar*** has been carried out under our supervision and that this work has not been submitted elsewhere for a degree.

It is further certified that the student has fulfilled all the requirements for the degree of Master of Technology.

Dr. Satish Kumar Sinha
Thesis Supervisor
Associate Professor
Department of Petroleum Engineering
and Geoengineering
RGIPT, Jais, Amethi

Dr. Alok Kumar Singh
Co-Supervisor
Associate Professor,
Department of Petroleum Engineering
and Geoengineering
RGIPT, Jais, Amethi

DECLARATION BY THE CANDIDATE

I, ***Vikram Kumar***, certify that the work embodied in this thesis is my own bonafide work and carried out by me under the supervision of ***Dr. Satish Kumar Sinha*** and ***Dr. Alok Kumar Singh*** from ***September 2020*** *to* ***July 2021*** at the **Department of *Petroleum Engineering and Geoengineering***, **Rajiv Gandhi Institute of Petroleum Technology, Jais (India).** The matter embodied in this thesis has not been submitted for the award of any other degree. I declare that I have faithfully acknowledged and given credits to the research workers wherever their works have been cited in my work in this thesis. I further declare that I have not willfully copied any other's work, paragraphs, text, data, results, etc., reported in journals, books, magazines, reports dissertations, thesis, etc., or available at websites and have not included them in this thesis and have not cited as my own work.

Date:

Vikram Kumar

Roll No: MPE19-005

Place:

CERTIFICATE BY THE SUPERVISOR

It is certified that the above statement made by the student is correct to the best of our knowledge.

Dr. Satish Kumar Sinha

Thesis Supervisor

Associate Professor

Department of Petroleum Engineering
and Geoengineering

RGIPT, Jais, Amethi

Dr. Alok Kumar Singh

Co-Supervisor

Associate Professor

Department of Petroleum Engineering
and Geoengineering

RGIPT, Jais, Amethi

Dr. Satish Kumar Sinha

Head of Department

Department of Petroleum Engineering and Geoengineering

RGIPT, Jais, Amethi

COPYRIGHT TRANSFER CERTIFICATE

Title of the Thesis: Machine Learning Approach for Vertical Permeability Prediction and Permeability Anisotropy of Hugin Sandstone Formation (Volve field).

Name of the Student: Vikram Kumar

Copyright Transfer

The undersigned hereby assigns to the Rajiv Gandhi Institute of Petroleum Technology, Jais all rights under copyright that may exist in and for the above thesis submitted for the award of the "Master of Technology".

Date:

Place:

Vikram Kumar

Roll No: MPE19-005

Note: However, the author may reproduce or authorize others to reproduce material extracted verbatim from the thesis or derivative of the thesis for author's personal use provided that the source and the Institute's copyright notice are indicated.

ACKNOWLEDGEMENT

I take this opportunity to express my deep sense of gratitude to my thesis supervisor Dr. Satish Kumar Sinha, Associate Professor and Head, Department of Petroleum Engineering and Geoengineering, Rajiv Gandhi Institute of Petroleum Technology (RGIPT), Jais, Amethi for his supervision, guidance, and meticulous examination of the thesis report. Without his help, it would not have been possible for me to carry out the work smoothly. I thank him for his valuable suggestion supervision and constant inspiration throughout this work and I am deeply indebted to him.

I am very thankful to thank Dr. Alok Kr. Singh, Associate Professor, Department of Petroleum Engineering and Geoengineering. I thank him for his valuable guidance and constant inspiration throughout this work. I also thank him for his periodic suggestion and morale boosting sessions.

I am grateful to my seniors Mr. Abul Aas, Mr. Krishna Raghav Chaturvedi, Mr. Alok Kumar, Mr. Himanshu Kesarwani, Mr. Ravi Shankar Kumar, Mr. Saurabh Tewari, Mr. Anurag Pandey, and Mr. Jitendra Argal for guiding me throughout the study and the rich technical discussions I had with them.

Special thanks to all faculty members of the Department of Petroleum Engineering and Geoengineering, RGIPT who contributed directly or indirectly to the accomplishments of this work and for their constant support and kind-heartedness.

I would also like to express my sincere gratitude to my family and colleagues Mr. Anjanay Pandey, Mr. Rishab Tripathi, Mr. Vishal Dhar and Mr. Romy Agrawal for their constant support and encouragement

I am thankful to Equnior (Statoil) for disclosing oil and gas filed data and making it open-source to support learning, innovation, and new solution to the energy future, and the possibility to extend the life of the field.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

mD:	milli Darcy
ML:	Machine learning
MLR:	Multiple linear regression
OLS:	Ordinary least square
ANN:	Artificial neural network
DT:	Decision Tree Regressor
GB:	Gradient Boosting Regressor
RF:	Random Forest Regressor
OLS:	Ordinary Least Square Regression.
IF:	Isolation Forest
PHIF, ø:	Porosity
SW:	Water Saturation
K_v:	Vertical Permeability
K_h:	Horizontal permeability
VSH:	Volume of Shale
HMR:	Hydraulic Mean Radius
BVW:	Bulk Volume Water
R_w:	Formation water resistivity at formation temperature

R_t:	True resistivity measured in formation
NMR:	Nuclear magnetic resonance
m:	Cementation exponent
n:	Saturation exponent
a:	Archie constant
R^2:	coefficient of correlation
MSE:	Mean Squared Error
MAE:	Mean Absolute Error
RKB:	Rotary Kelly Bushing
GR_{max}	Clean sand gamma-ray value (API)
GR_{max}	Shale gamma-ray value (API)
GR_{log}	Gamma ray log reading (API)

PREFACE

In the desertion to understand the anisotropic nature of well #1 (Hugin formation of Volve field). The measured depth (RKB) interval of the Hugin formation top 3821.5 m to 3919.6 m. The prediction of vertical permeability for well #1 at the same depth interval is carried throughout the thesis. There are two different data set are considered to train five machine learning models. The first dataset case 1 considered as vertical permeability of core data as target and corresponding well logs feature is used to train all the considered machine learning models. In case 2 the considered training data is two well logs they are well #2 and well #3, the target vertical permeability estimated by Statoil is available for both well and the features of the well logs are considered the same as case 1. However, for this study, the concept of featuring engineering and domain is used and introduces hydraulic mean radius as a training feature for both cases. The training data must be per-processed before using it to train different models. The data processing is done in consideration to not lose the properties and originality of well logs data. The few techniques are used for data processing, are data cleaning, feature engineering, logs smoothening, normalization, outliers removal. The five ML models used regression type supervised machine learning models; two of them are shallow models are multi linear regression and artificial neural network, and three of them are tree-based deep learning models based on decision tree regression, gradient boosting regression and random forest regression. Performances of all ML models have been evaluated based on three metric scores, namely, coefficient of correlation (R^2), mean squared error, and mean absolute error. The Random Forest model outperformed other models in both cases. We ensure the stability of the random forest model by cross-validation technique before predicting vertical permeability. The predicted permeability is validated with the core data available. The investigation of permeability anisotropic of entire Hugin formation and found the formation is vertical anisotropic in nature. This study computed the correlation from the predicted vertical permeability to function of porosity and horizontal permeability was derived by using (OLS) Ordinary least square regression.

CHAPTER – 1: INTRODUCTION

1.1 Background

The exploitation of oil reservoirs with maximum oil recovery needs precise determination of petrophysical parameters such as porosity, permeability, water saturation, etc. (Aliouane et al., 2012). They play a key role in quantitatively characterizing subsurface reservoirs and aid us in estimating Original Oil in Place (OOIP) and permeability anisotropy. Their role is also critical in developing a reservoir management plan, drilling a new development well, optimization of completion design, and enhanced oil recovery (EOR) injection strategy (Aliouane et al., 2012; Barakat & Nooh, 2017). Among them, permeability is one of the key parameters that govern the in-situ fluid flow and thus dictates oil recovery results. It is, in simplest terms, is the ability of the rock to transmit fluids through it. Also, two types of permeability depending upon the direction are generally Aliouane (2012) discussed in the literature (Zagrebelnyy et al., 2017). Vertical permeability is perpendicular whereas horizontal permeability is parallel to the bedding plane. It is noted that horizontal permeability (K_h) is generally greater than vertical permeability (K_v), especially, when sand grains constituting reservoir rock are irregular and small. Most of the petroleum reservoirs around the world fall in this category. The ratio (K_v/K_h) is important as it represents the contrast between horizontal and vertical planes within the subsurface formation. This, in particular, aid in reservoir simulation studies where its precise input can effectively simulate future recovery, artificial lift selection, optimal drainage points and production rate, etc. In layered stratigraphic reservoirs, each layer has different vertical permeability, and thus different (K_v/K_h) relationships (Hou et al., 2016; Zagrebelnyy et al., 2017).

Various conventional techniques that are considered standard in the petroleum industry are used for permeability determination such as coring analysis, well log interpretation and well test analysis. Several studies can be found in the literature incorporating different techniques for permeability determination for a variety of reservoirs. Zahaf & Tiab (2000) describe vertical permeability as a function of K_h, hydraulic mean radius, water saturation and develop the model for estimating in situ vertical permeability for TAGI formation. Meyer (2002) describes the permeability anisotropy of Viking formation sandstone by using core samples with probe permeameter gas flow measurement and found that it varies with depth. Another group of researchers McCabe & Horne (2015) describe the permeability anisotropic from the distribution of subsurface temperature by modeling the temperature distribution from DST tools. Further Sheng (2008) incorporated the use of single probe tests in horizontal well and found that both vertical and horizontal permeability can be estimated. Shedid (2019) derived the different empirical correlations of K_v and establish the anisotropic ratio from core data. J. Wang et al., (2020) used X-ray computed tomography and found the degree of anisotropic of hydrate sediments. Also, a comprehensive approach is adopted in some studies by integrating well log and core data for better reservoir description. As a result, various correlations were developed for determining petrophysical parameters and their accuracy was tested with established techniques. Johnson (1994) estimated permeability by identifying different hydraulic zone using well logs, available core data, and wireline formation tester data and found that the response of permeability is different for the different hydraulic zone. Anderson (1994) found that the resistivity logs were able to capture the pattern of fluid invasion in highly permeable beds and stated that this movement of invaded fluid was used for estimating formation permeability. One study direct measured local permeability and permeability anisotropy is given by Ayan et al., (2007) using a formation tester incorporating pressure transient testing data as well as a sonic image log. Moreover recently, the use of machine learning techniques gained much interest among researchers which is motivated by its application for petrophysical properties determination with robust models. From the last two decades, the use of the data-driven model for subsurface and pytrophysical study increased for the last year more than 1500 research article including more than 25 research article (well logs plus machine learning) published. Mohaghegh et al., (1997) introduced artificial neural network for multi variable regression to predict permeability from the responses of well log and compared with different empirical correlation. Lim (2005) predicted permeability using fuzzy logic and neural network and describe the reservoir properties in spatial variation using well logs. Finally, he concluded that intelligent techniques can be powerful tools for the determination of reservoir properties. Salem et al., (2019) estimated permeability values by well logs using neural network using geological diagenesis as a feature and thereby capable of predicting porosity and permeability of carbonate reservoir. Valent (2019) used deep autoencoders deep network technique from borehole image logs data and support vector machine to estimate permeability log. Thus, different ML algorithms gather the need to be developed for the comprehensive use of data-driven modeling in the determination of petrophysical properties.

Hence in this work, there are two objectives both are related each other the first objective is to predict the high-resolution K_v and validate with core available core data and another objective was to drive empirical correlation using predicted we used full well logs and available core data to study the reservoir anisotropic of Hugin sandstone formation of Volve field. To study reservoir anisotropic, we used the predicted high-resolution K_v thus the intelligence technique required to develop a data-driven model for estimating $K_{v.}$

Better understanding the data-driven modeling better the prediction and helps in model development, the workflow has implemented a bunch of techniques required to construct and develop model these are data pre-processing (Nordloh et al., 2020; Singh et al., 2020) data smoothing (Chaki et al., 2018), treating outliers (Li & Misra, 2021), normalization (Li & Misra, 2021), machine learning models (Nordloh et al., 2020), model evaluation, model stability investigating (Li & Misra, 2021), etc. used workflow makes the data-driven model robust as shown in figure 4.1 & 4.2. The machine learning models are used for training for five regression types supervised machine learning model having 2 shallow models are multi linear regression (MLR) and artificial neural network (ANN), and three tree-based models they are decision tree regression (DT), gradient boosting regression (GB) and random forest regression (RF). The best models were investigated among five models for the prediction of K_v. The best model is used for two different approaches to training data. The first approach of ML models training with two well logs (well #2 & well #3) where the dependent feature or target is K_v and other independent features of well logs are porosity, water saturation (SW), volume of shale (VSH), bulk volume water (BVW), hydraulic mean radius (HMR), horizontal permeability (K_h). Except for HMR, these independent features are estimated by Statoil mention in the well report. The HMR is a new feature derived from domain knowledge and feature engineering concept, it is used for training the models. The second approach of prediction is to train the different ML models algorithms considering core data K_v as target/dependent feature and the independent variable corresponding to well logs (well #1). A modern approach to training the models with experimental data, however, the training data samples are limited to 86 after outlier's removal.

However, the best model is used to evaluate from the three performance metrics all models are the coefficient of correlation (R^2), mean squared error (MSE), and mean absolute error (MAE). The cross-validation technique is used to best model to test and ensure the model stability. The considered empirical correlation model for K_v of sandstone cross plot against Hugin sandstone core data (Shedid, 2019). The different cross-plot of prediction of both approaches of training models is in figure 6.5. Thus, we have developed a correlation for Hugin sandstone formation of

K_v, K_h, and porosity can be moreover, the developed correlation for reservoir anisotropic of Volve field.

1.2 Field Overview

Equinor in Oct-Sept 2018 a new Version of file set was released by replacing the beta version of file set is Open Data disclosed under a License and to protect the rights of the data owner. The massive dataset is Stored in the Microsoft Azure database.

The Volve field is located 200 km from Stavanger in the North Sea as shown in figure 1.1. The Volve field is described as a fault block structure with an initial estimation of 173 million bbl of oil in place (https://data.equinor.com/dataset/Volve). The Volve dataset consists of approximately 40,000 files of Exploration and Production. There is no known aquifer support, so the drainage was primarily dependent on reservoir depressurization and hydrocarbon displacement by water injection. Hence there are certain production and injection well data are available for the field. The Volve field was decommissioned in September 2016 after roughly nine years of production and operation that delivered 63 million barrels, with a recovery rate of

54 % (Sen & Ganguli, 2019; B. Wang et al., 2021). The intention of disclosing the bulk dataset is to support learning, innovation, and new solution to the energy future, and the possibility to extend the life of the field.

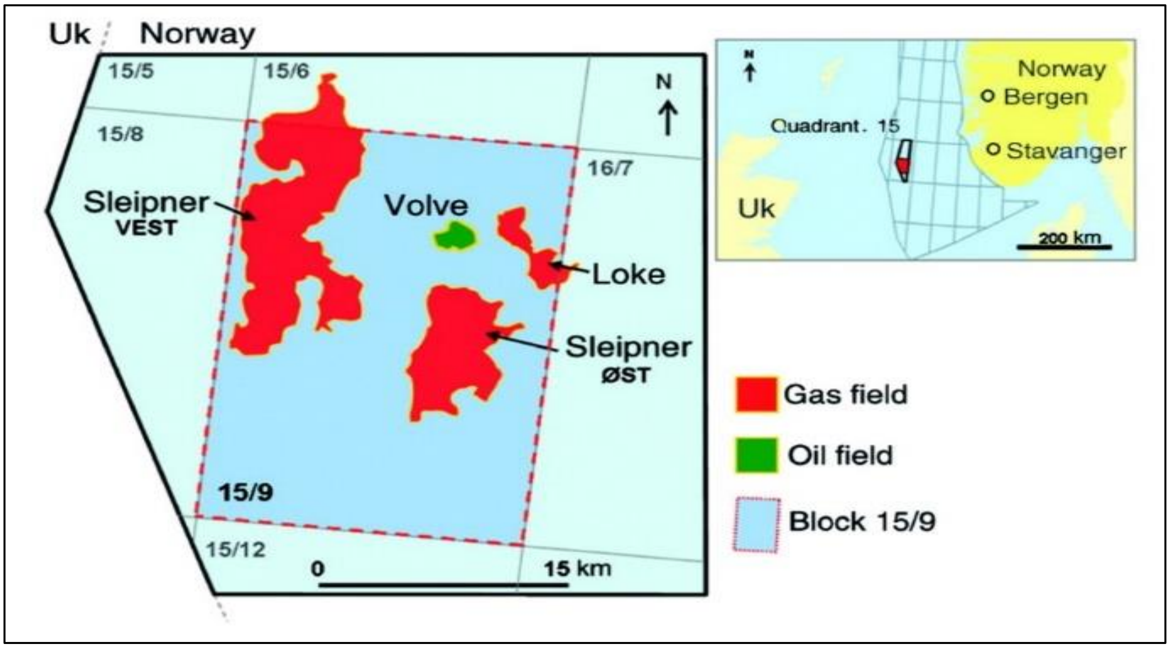


Figure 1. 1: Location of Volve field in the North Sea (adapted from Ravasi et al. 2015).

CHAPTER-2: DATA DESCRIPTION

2.1 Data considerations for this study

In the petrophysical part of the Volve dataset, there are 24 well data given of this fields among some wells are having some information of core data which is experimentally based and having considerable data samples of core was available for (15/9-19 A), hence we can use this core data for validation. The selected wells for this study 15/9-19 A, 15/9-19 BT2, and 15/9-19 S&SR are named as Well #1, Well #2, and Well #3 respectively.

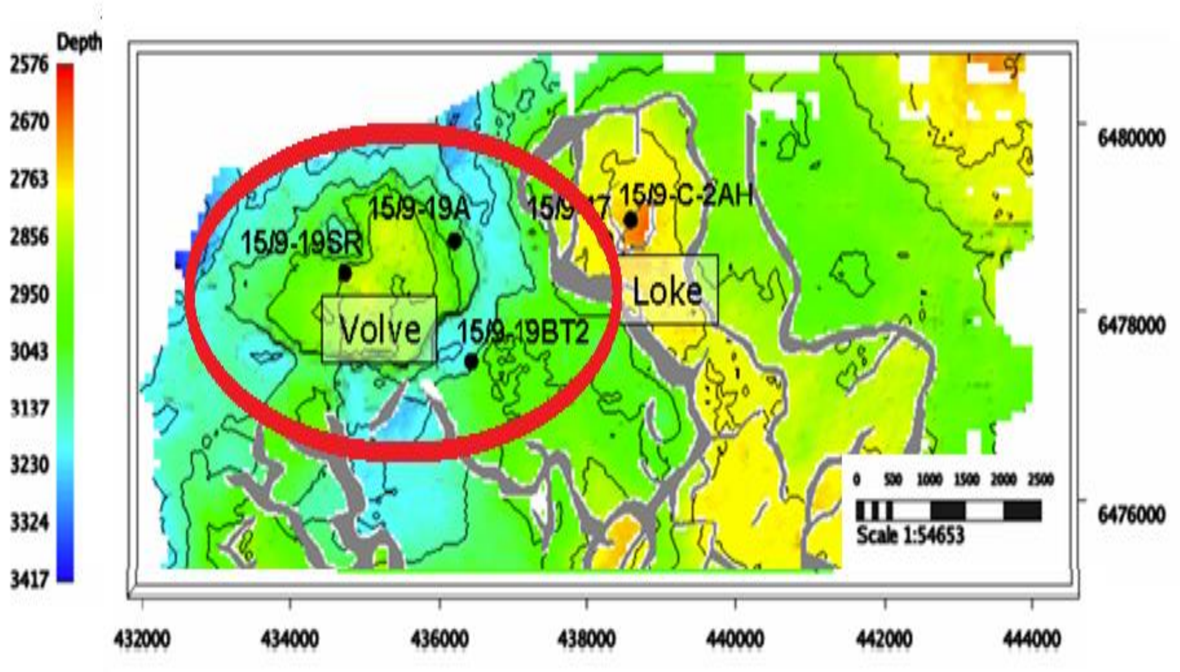


Figure 2. 1: Location of considered wells for study shows heat map as the depth of tops of Hugin formation (courtesy Statoil).

2.2 Data quality and handling missing data

The quality of data can highly influence the ML model negatively and restrict the development model or it may be unusable. The considered data for this study well and good, however, some sections of the well logs data are missing as mentioned in figure 2.2. The reason for abnormal data while measurement of log the formation signal, random noise or logging tool malfunction sensor error (Akkurt et al., 2019). The data which we are dealing with has an imbalance in distribution that has to be processed before feed as input for ML models. The data processing techniques are based on data visualizations and exploratory data analysis. Due to the high variance in some of the logs that are (standard deviation upon mean) hence the smoothing techniques are applied for logs having high variance shown in table 3.2 shows list of logs with variance.

Log intervals are selected based on the availability of data, a large interval of gap or missing data should be avoided for good prediction, however, some of the sample data points missing values are represented as (-999) NAN indicated non-numeric numbers are re-filled the missing data with mean or average values of above and below sample data points shows in figure 2.3.

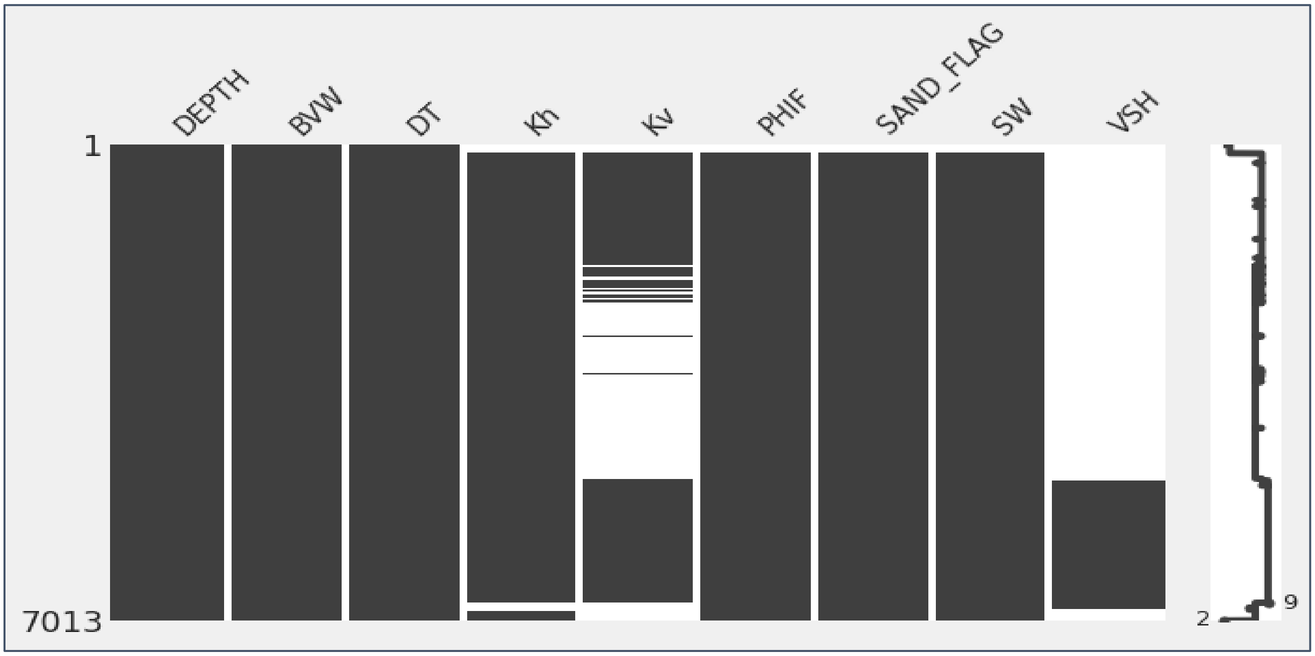


Figure 2. 2: Visualizations of missing data of well #3, in the logs white part, represents missing data, and black represents data available.

On the most right figure, 2.2 is a sparkline that ranges from 0 on the left to the total number of columns in the data frame on the right. A closeup can be seen above. When a row has a value in each column, the line will be at the maximum right position. As missing values start to increase within that row the line will move towards the left. Therefore the above sparkline at most below shows 2 logs have no missing data by interpretation the logs may be depth and BVW. The sparkline shows nine up to some interval this means all 9 logs don't have missing/NAN values in well #3.

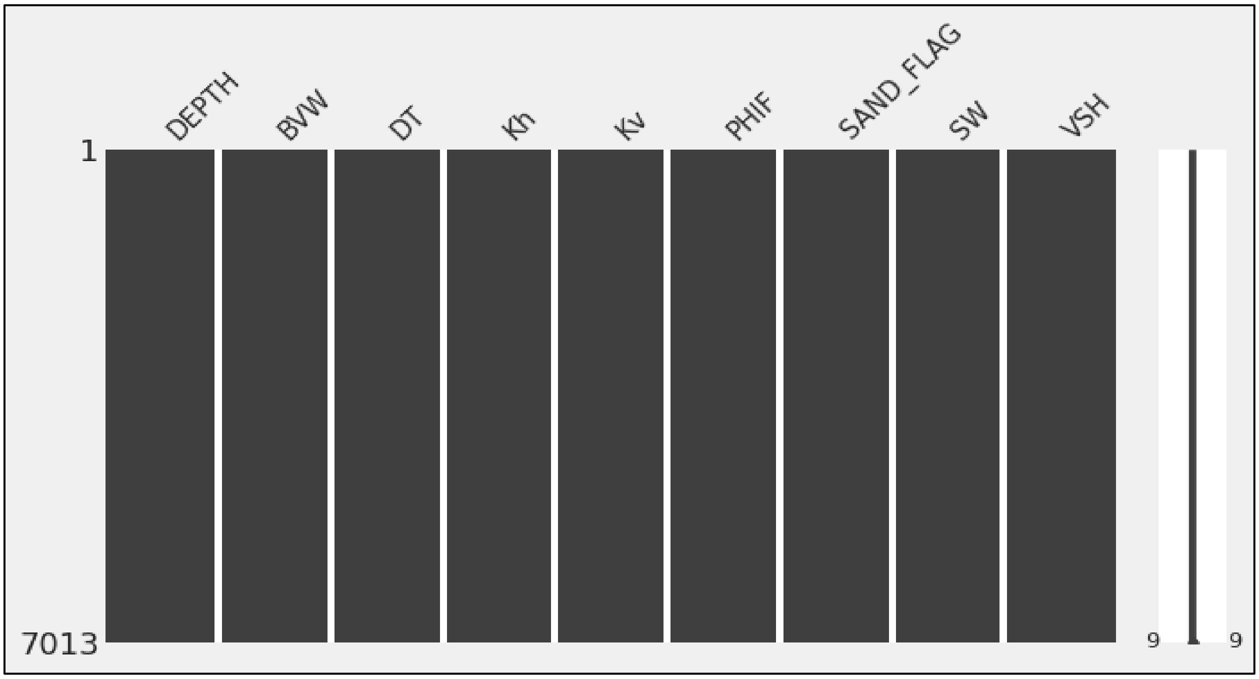


Figure 2. 3: Visualizations of well #3, after filling the missing values with mean values of above and below sample data points.

2.3 Data concern and model input attributes

Each of these wells has what they are known as Logs. The logs are certain physical measurements that represent the properties of each rock strata over the depth. The well logs of each concern wells have attributes is computed from data which is measured from the sensor/tools except K_h and K_v. There is no well log which can directly determine permeability however, the nuclear magnetic resonance (NMR) log measures permeability in terms of fluid distribution.

The following is a list of logs that can be used for model input.

- PHIF is the formation density porosity in v/v.
- SW is the water saturation in the fraction v/v.
- VSH volume of the shale is calculated as gamma Minimum (sand indication) and gamma maximum (shale indication).
- BVW is the bulk volume of the water is the product of water saturation and porosity. BVW is an indicator of hydrocarbon and reservoir homogeneity (Mabrouk, 2005).
- K_h obtains from multivariable regression analysis along with corrected with reservoir condition in millidarcy (mD).
- HMR hydraulic mean radius is obtained from the square root of liquid permeability upon porosity (mD).
- K_v is the vertical permeability (mD).

The K_h and K_v are soft data, computed from multivariable regression by Equinor by using well logs against overburden corrected core permeability (Volve petrophysical report, 2014). Moreover, we used K_h and K_v for model training however, if we train our models with wireline NMR, K_h and K_v accurate measured data no doubt the model will robust (Rosenbrand et al., 2015). The use of soft data (estimated/computed) for our models for K_v prediction is robust because we used enough models training samples and ensure the stability of our model including model validation with available core data. The model training from hard data (measured data) is always a good option for model training. Therefore, again we trained our model with core K_h and K_v corresponding to well logs. We found the model was fine with good accuracy over test data, however, we compare the results in the upcoming section.

CHAPTER-3: DATA PROCESSING

3.1 Feature engineering

Feature engineering is the process of using your domain knowledge about the data and the machine-learning algorithms at hand to make the algorithm work better by applying hardcoded transformations to the data before it goes to the machine learning model. The inputs/features in our data will directly influence the ML models, to obtain a useful model which can be implemented for field development. For similar purposes, with well-engineered features, can be a very simple model and have good results. We don't have to put in as much effort to choose the best models and parameters. Therefore, we introduce hydraulic mean radius (HMR). The other feature is estimated from measured data they are SW, VSH, PHIF, BVW, $K_{h,}$ and K_v. The information is available in the Volve field report (Volve petrophysical report, 2014).

The other feature is estimated as bulk volume of water (BVW) is simply the product of water saturation (SW) and porosity (PHIF). The true BVW shows reservoir is whether or not is at irreducible water saturation (Asquith, 1985). This is the most important parameter to indicate the hydrocarbon potential zone (Mabrouk, 2005).

The continuous log horizontal permeability (K_h) is estimated based on multivariable regression analysis between well logs and porosity (PHIF) and shale volume (VSH) (normalized gamma-ray log) against overburden corrected core permeability (Volve petrophysical report, 2014). The equation was used.

$$K_h = 10^{(2+32.PHIF-9.VSH)} \tag{3.1}$$

The SW is calculated using the Archie equation,

$$S_w^n = \left(\frac{a.R_w}{R_t.\emptyset^m}\right) \tag{3.2}$$

Where,

Sw = Water saturation (fraction)

a = Archie constant (= 1.0)

Rw = Formation water resistivity at formation temperature (Ohm.m)

Rt = True resistivity measured in formation (Ohm.m)

φ = Porosity (fraction)

m = Cementation exponent

n = Saturation exponent.

The VSH is computed from the gamma radioactivity from minerals in petrophysical logging is measured on the API scale. The most common gamma-emitting lithology is shale. This is because shales are ultimately derived from igneous rocks which have significant amounts of gamma-emitting isotopes, they are potassium and occasionally also U-Ra and Th-series isotopes. Potassium beds are also highly Radioactive and measure high gamma reading, whereas the sandstone dominated with quartz measures low gamma radiation (Basin, 2017). The shale volume is derived from the gamma-ray maximum, GR_{log}, using a linear relationship, as shown in the below equation.

$$VSH = \frac{GR_{max} - GR_{log}}{GR_{max} - GR_{min}} \tag{3.3}$$

Where,

GR_{max} = Clean sand gamma-ray value (API)

GR_{max} = Shale gamma-ray value (API)

GR_{log} = Gamma-ray log reading (API)

Table 3. 1: The list of gamma minimum and maximum of all three wells, (courtesy Statoil).

Well	**Area**	**Hugin formation**	
		GRmin(API)	**GRmax(API)**
15/9-19A/ Well #1	Volve	12	115
15/9-19BT2/ Well #2	Volve	8	100
15/9-19SR/ Well #3	Volve	17	135

The HMR is one of the most important properties that determine how much fluid a channel can discharge and how well it can move sediments. A large hydraulic radius value means a smaller volume of contact fluid and a larger cross-sectional area in the channel. The hydraulic mean radius (HMR) considers variations in permeability over porosity as it is defined to be equal to root over liquid permeability upon porosity (Shedid, 2019). It is very important to understand a relationship between microscopic level attributes and microscopic core data based on the concept of HMR.

3.2 Smoothing technique

Abnormal quality of data can highly influence the ML model negatively and restrict the development model or it may be unusable. The reason for abnormal data while measurement of well log the formation signal, random noise or logging tool malfunction /sensor error. (Akkurt et al., 2019). Therefore, the considered training dataset had high variance (standard deviation upon mean), this variance may due to high deviation within layers of the formation. The smoothing approach can good influence to be processed before a feed as input for ML models.

Table 3. 2: Shows computed high variance using standard deviation and mean.

Index	Well Logs	Standard Deviation	Mean	Variance
0	PHIF	0.06	0.18	0.26
1	SW	0.11	0.96	0.11
2	VSH	0.14	0.28	0.5
3	BVW	0.06	0.22	0.27
4	HMR	66315	19878	1.45
5	K_h	22504	6322	3.56
6	K_v	26510	6982	3.8

As we look at the statistics of training data in table 3.2, high variance logs are present in our training dataset these are 'K_v', K_h, HMR. The logs must be smoothing by moving average filter, however in literature, there are many other filters are used for log smoothing (Duchesne & Gaillot) (2011) used, Three types of filters among all discrete wavelet transform is the best filter for acoustic logs however the binomial filter and average filter are well suited for permeability logs smoothing, as we can see the reduced variance in above table 3.3. There are many other filters used for better smoothing for different logs(Chaki et al., 2018).

This study used a simple moving average filter which is a loss pass filter and is often used for an array of sampled data or signals that helps to filter undesirable noisy components from our training data(Chen & Chen, n.d.). We consider 9 samples of input at a time for Kv and Kh and compute the average of those 9 samples and generates new output data with low variance. However, for HMR we consider 10 samples points at a time. Because as sample points of input increase, the more smoothed output we get. Therefore, we can observe the filter HMR intended smoother over true HMR.

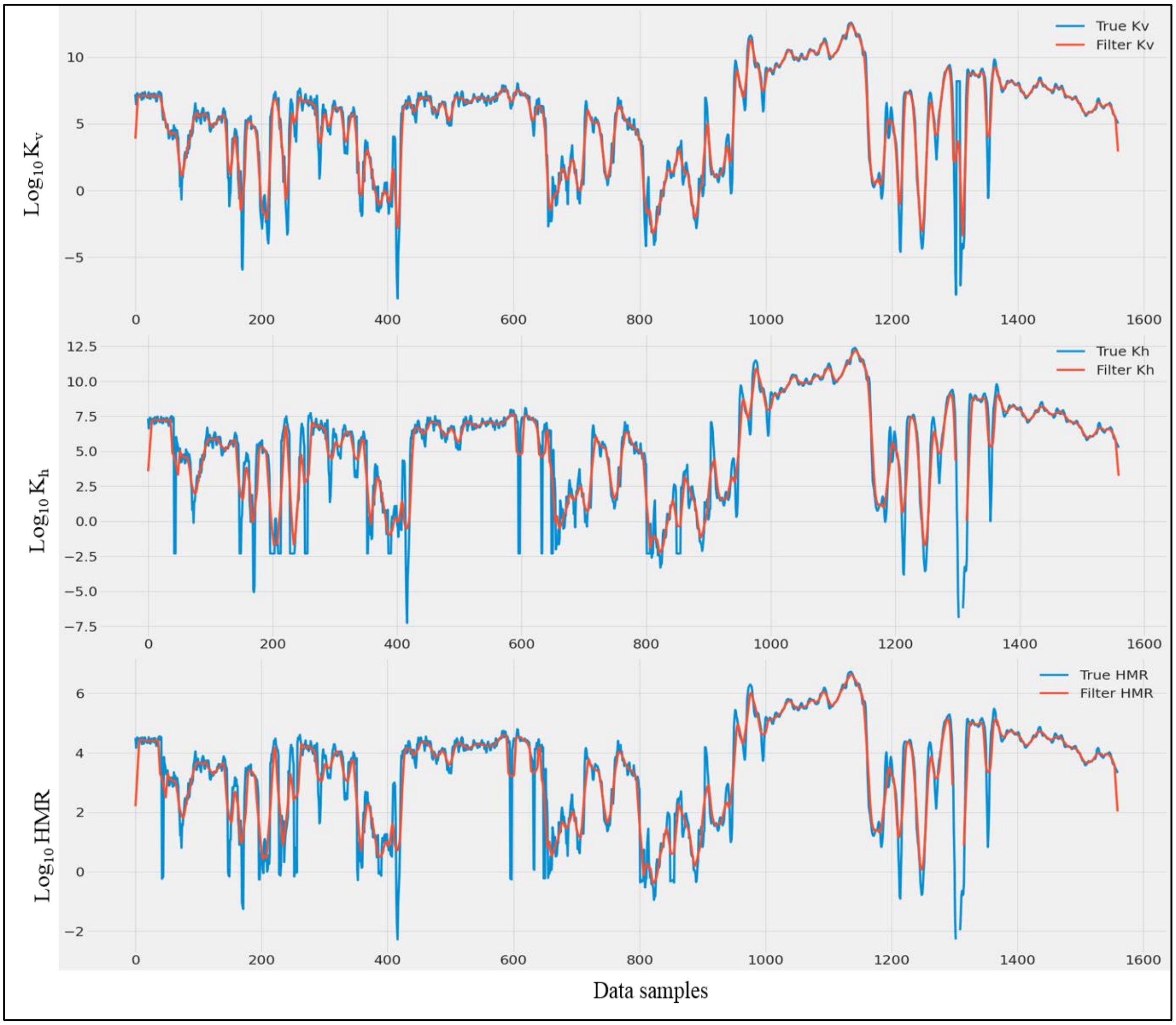


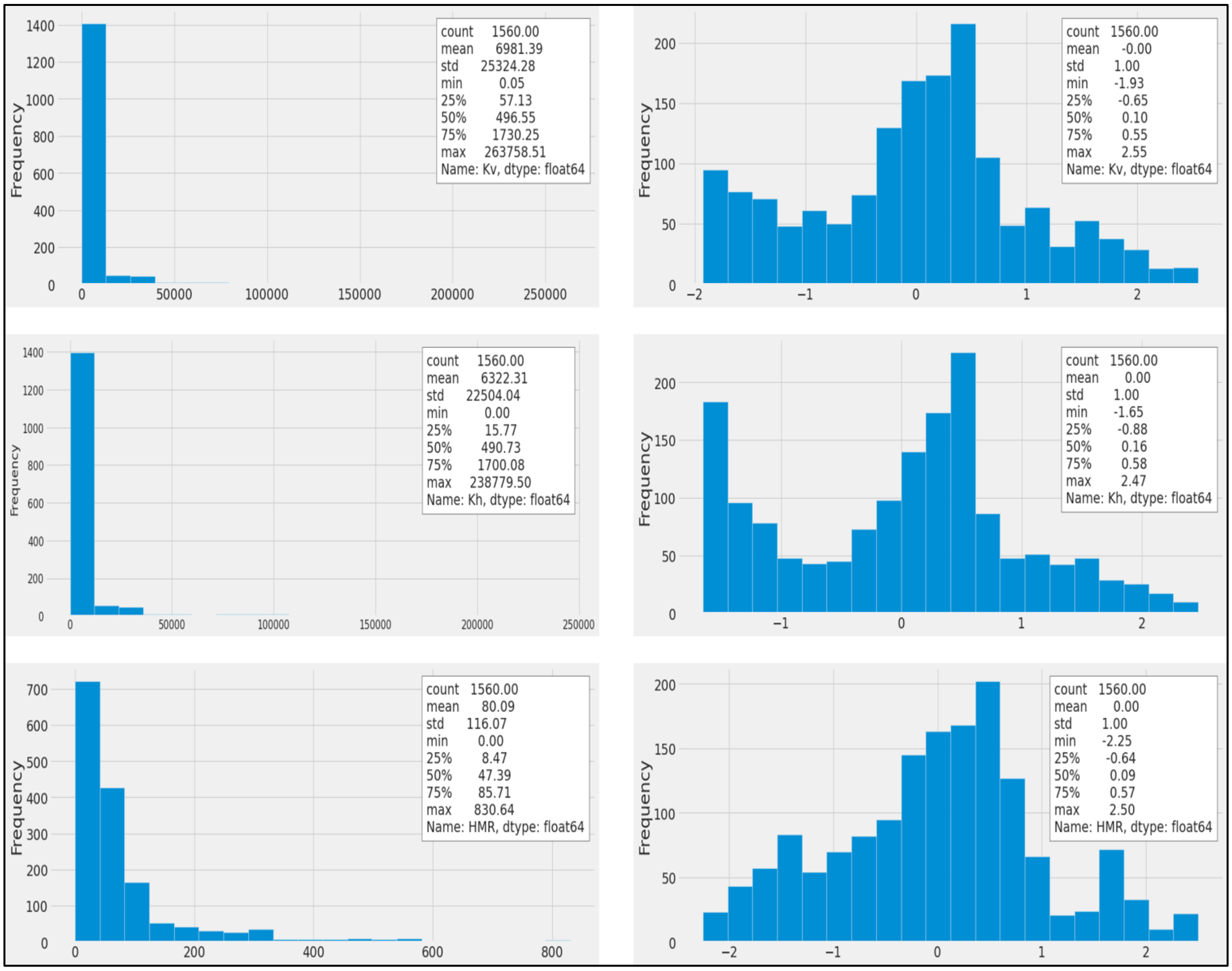
Figure 3. 1: The implemented moving filter to high variance data is Kv, Kh, and HMR over sample points 1560. The blue curve shows true training data and the red curve shows smoothed training data.

Table 3. 3: Shows computed reduced variance using standard deviation and mean smooth training data.

Index	Well Logs	Standard Deviation	Mean	Reduced variance	Old Variance
0	PHIF	0.06	0.18	0.26	0.26
1	SW	0.11	0.96	0.11	0.11
2	VSH	0.14	0.28	0.5	0.5
3	BVW	0.06	0.22	0.27	0.27
4	HMR	108	80	1.35	1.45
5	K_h	556	398	1.4	3.56
6	K_v	536	354	1.51	3.8

3.3 Normalization

A dataset usually contains features that are considerably different in magnitude, measurement unit, and range from one another. Example in our training data notices the range of water saturation values lies between 0 to 1 and the range of K_v is 0.05 to 263758 mD. The ML approaches are based on Euclidean distance, gradient, density, and volume (Li & Misra, 2021). Normalization of the dataset can monotonic transformation of the data into Gaussian-like distribution without changing the data and helps them being treated equally in the eyes of ML laws (Li & Misra, 2021; Sola & Sevilla, 1997). We implemented power transformer Yeo-johnson methods to stabilize variance by considering the standard deviation as 1 and minimize skewness that can help to achieve Gaussian distribution. Therefore, more Gaussian distribution of training data more suitable input data for the ML model. Before and after normalization of the processed training data look like 1560 training samples in figure 3.2.



```
count     1560.00
mean      6981.39
std      25324.28
min          0.05
25%         57.13
50%        496.55
75%       1730.25
max     263758.51
Name: Kv, dtype: float64
```

```
count   1560.00
mean      -0.00
std        1.00
min       -1.93
25%       -0.65
50%        0.10
75%        0.55
max        2.55
Name: Kv, dtype: float64
```

```
count     1560.00
mean      6322.31
std      22504.04
min          0.00
25%         15.77
50%        490.73
75%       1700.08
max     238779.50
Name: Kh, dtype: float64
```

```
count   1560.00
mean       0.00
std        1.00
min       -1.65
25%       -0.88
50%        0.16
75%        0.58
max        2.47
Name: Kh, dtype: float64
```

```
count   1560.00
mean      80.09
std      116.07
min        0.00
25%        8.47
50%       47.39
75%       85.71
max      830.64
Name: HMR, dtype: float64
```

```
count   1560.00
mean       0.00
std        1.00
min       -2.25
25%       -0.64
50%        0.09
75%        0.57
max        2.50
Name: HMR, dtype: float64
```

Figure 3.2: High variance training well logs attributes distribution before and after normalization.

3.4 Outliers removal

We observe out-of-range (high standard deviation) data that are outliers in the box and whisker plot figure 3.3. The outliers are present in our data set due to many reasons as we discussed above. Outliers were present in our dataset the ML models maybe not be robust as well stable as after removing extreme values.

Standard deviation and isolation forest techniques are used to detect and discard the outliers. In standard deviation methods, we need to set a desire threshold value as the minimum and maximum away from the standard deviation technique to keep the data in between min and max, and the rest of the data are considered as outliers. However standard deviation method maybe not be an effective way to discarding outliers from the well logs training dataset by visualization box and whisker plot in the training dataset before and after outliers were removed.

The other method is used to removed Outliers removed using an isolation forest. Isolation Forest (IF) assumes that outliers are more likely to be found in sparse parts of the feature space, with more vacant space surrounding them, than thickly clustered normal/inlier data. Because outliers are found in sparsely populated areas of the dataset, isolating them in a segment/partition usually requires fewer random partitions. By randomly subsampling features and their related threshold values, IF achieves recursive random partitioning/splitting of the feature space. The number of splits necessary to isolate a sample at a terminating node is equal to the path length from the root node to the terminating node, resulting in a treelike structure. This path length, averaging nearly over a forest of all such random trees, is a measure of a sample's normality, such that anomalies or outliers have noticeably shorter path lengths; or simply it is easy to partition the outliers with a small number of partitions of the feature space. In the training dataset, 155 samples are deleted from the well logs based on the assumption that 0.09 as contamination factor in the training data points are outliers and left with 1405 training samples. There are a total of 1560 training samples present, the standard deviation method was able to remove only 7 training samples as outliers and left 1553 training samples. However, the standard deviation method in figure 3.4 is only able to discard the extreme values from the training sample, moreover, the IF methods can remove outliers effectively as we can the box and whisker in figure 3.5. However, very few outliers are shown in BVW and one outlier in K_v log. The data processing needs to be done for usable ML models however at the same time we have to preserve the quality of original data.

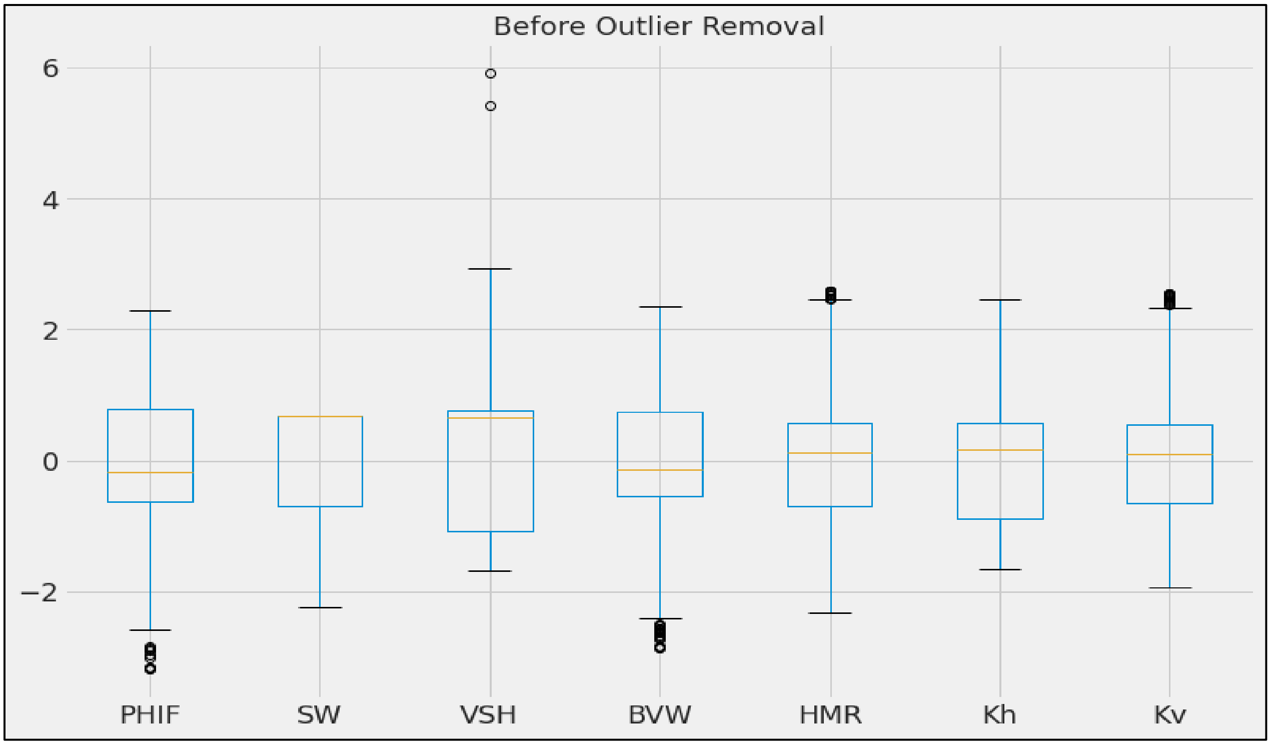


Figure 3.3: The box and whisker plot of raw training samples before outliers removal.

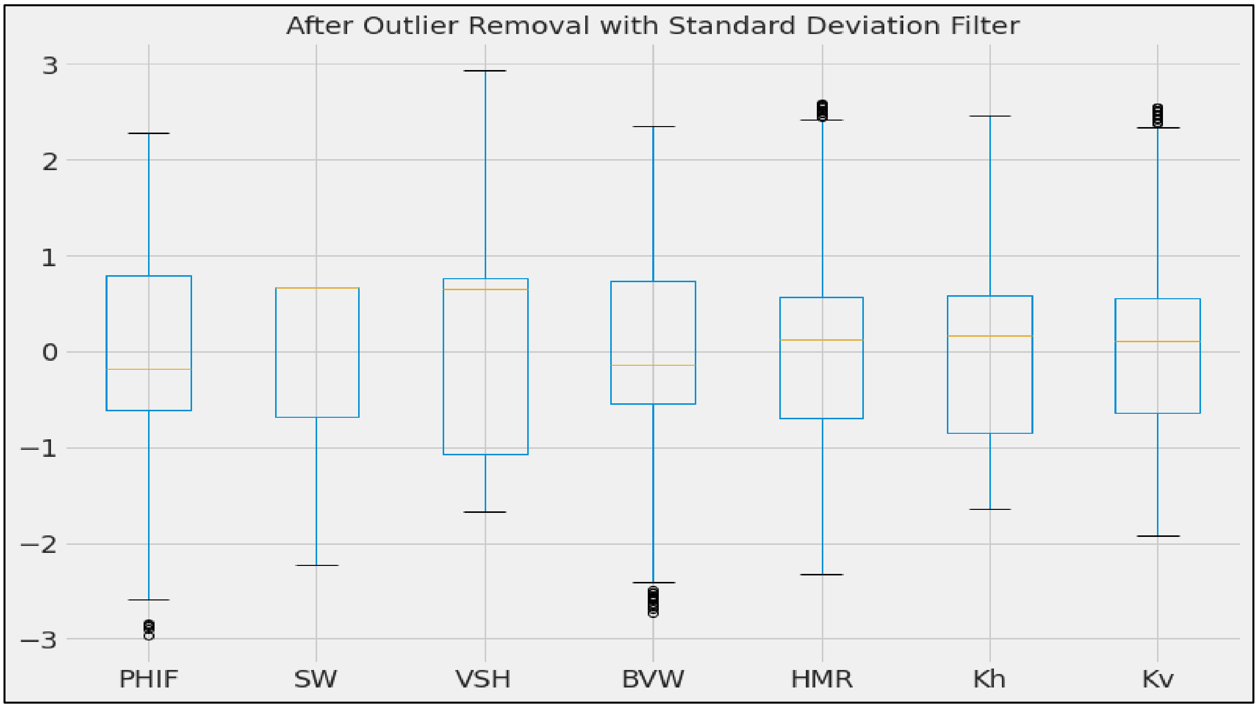


Figure 3.4: The box and whisker plot of training samples after treating with standard deviation outliers removal methods still showing outliers in some logs.

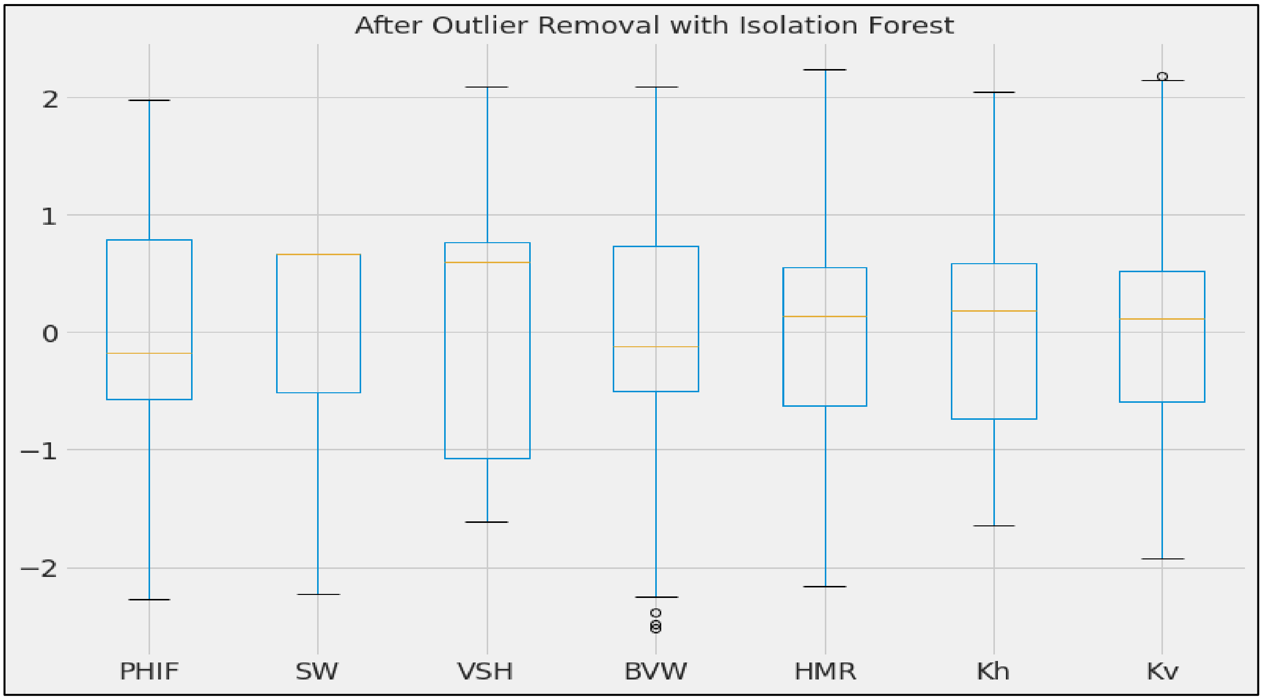


Figure 3.5: The box and whisker plot of training samples after treating with isolation forest outliers removal methods and removed maximum outliers as 155 sample points from the training sample.

3.5 Feature correlations

After treating the training dataset with outliers' removal methods, the correlations among well logs are calculated to check the correlation coefficient of input well logs features with extreme low correlation index can impact the ML model significantly. Using domine knowledge all input features are directly or indirectly related to the target, further, we ensure by calculating two different methods are used to compute correlation coefficient, One of the most frequently used correlation coefficients for measuring the linear correlation between two variables is the Pearson correlation coefficient. It is calculated by dividing the covariance of the two features by the product of their standard deviations, which is represented as

$$\rho(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2(Y_i - \bar{Y})^2}} \tag{3.4}$$

If X and Y denote the two features, n is the number of samples, and i denotes the sample index. The other Spearman correlation coefficient is used to determine whether two features have a monotonic relationship. The Spearman correlation coefficient measures both linear and nonlinear correlations, as opposed to the Pearson correlation coefficient. The Spearman correlation coefficient incorporates the Pearson correlation coefficient of two rank variables, which is expressed as

$$r_s = \rho_{rgx,rgy} = \frac{cov(rg_{X,}\, rg_Y)}{\sigma_{rg_{X,}}\, \sigma_{rg_Y}} \tag{3.5}$$

Where *cov* represents the covariance, $\boldsymbol{rg_X}$ and $\boldsymbol{rg_Y}$ denotes the rank of the two variables, and σ is the standard deviation. Tables 3.4 and Table 3.5 demonstrate the Pearson and Spearman correlations of the input logs (i.e. features) respectively. The K_v and K_h are high linear correlations that represent homogeneity (degree of homogeneity is high). The correlation values are 0.92 and 0.95 computed using Pearson correlation and Spearman correlation respectively, this is the maximum correlation computed. The porosity and K_h show high correlation index than porosity and K_v thus we can say that conductivity of porosity is more for parallel beds. However, the values of K_v are more than K_h which concludes the reservoir fall in horizontal anisotropy categories. The SW shows a low correlation index to other features not more than 0.26 we can discard SW features from input features, however, the domine knowledge says the SW is a good feature for both permeability and mark the fluid types in the reservoir as a marker for permeability and porosity shows Spearman correlation index for SW and porosity is -0.23 its satisfy Archie’s equation

(Mohamad & Hamada, (2017). Hence we keep the SW as input features. The VSH and both permeability show a high negative correlation of -0.82 and -0.75 for K_v and K_h respectively thus the shale tends to high compaction under overburden conditions, therefore, the porosity and permeability are very sensitive shale within the reservoir.

Table 3. 4: Pearson coefficient of the input/feature and target well logs.

	PHIF	SW	VSH	HMR	BVW	K_h	K_v
PHIF	1	-0.23	0.11	0.13	0.92	0.22	0.13
SW	-0.23	1	-0.26	0.26	0.06	0.22	0.24
VSH	0.11	-0.26	1	-0.88	0.03	-0.84	-0.94
HMR	0.13	0.26	-0.88	1	0.2	1	0.95
BVW	0.92	+-----0.06	0.03	0.2	1	0.29	0.21
K_h	0.22	0.22	-0.84	1	0.29	1	0.95
K_v	0.13	0.24	-0.94	0.95	0.21	0.95	1

Table 3. 5: Spearman correlation coefficient of the input/feature and target well logs.

	PHIF	SW	VSH	HMR	BVW	K_h	K_v
PHIF	1	-0.18	0.08	0.08	0.82	0.15	0.1
SW	-0.18	1	-0.2	0.2	0.05	0.17	0.18
VSH	0.08	-0.2	1	-0.79	0.02	-0.75	-0.82
HMR	0.08	0.2	-0.79	1	0.13	0.96	0.89
BVW	0.82	0.05	0.02	0.13	1	0.2	0.14
K_h	0.15	0.17	-0.75	0.96	0.2	1	0.92
K_v	0.1	0.18	-0.82	0.89	0.14	0.92	1

CHAPTER – 4: MACHINE LEARNING APPROACH

4.1 Data-Driven workflow

The data used for ML model for training, validation, and prediction contains three oil wells and core data of one well within the same reservoir. Our main purpose is to predict K_v and investigate the reservoir anisotropy of well #1 from high-resolution data (3820 meters to 3920 meters) 100 meters of logs contain 619 sample points. The prediction of K_v from two different training using different data. The predicted K_v is validated with core data of same well #1, available with depth interval (3837 meters to 3965 meters) within this interval there are 96 data samples available. Further develop a correlation between predicted K_v, $K_{h,}$ and porosity as well as drive correlation from predicted K_v to evaluate permeability anisotropy (K_v/K_h) between porosity and K_h.

For now, we aim to construct two different approaches to supervised machine learning algorithms are shallow learning model and the three tree-based models including two ensemble models used for training two different datasets by considering two cases. A complex and hybrid model with many parameters may not necessary and chances of overfitting for this low dimension regression problem. The feature/input and target/output (K_v) have low dimensionality and can simple relationship. The model was used for this study is supervised types machine learning algorithm. The shallow learning algorithm is constructed as relatively simple machine learning methods that computed around (10-80) parameters at the time of training the model. The shallow models are multi-linear regression (MLR) and shallow artificial neural networks (ANN) with 2 hidden layers including each layer that contained five neurons. The other type of model is based on the decision tree, The tree-based model can be shallow or not based on the number of training data samples trees and the number of split nodes. The single decision tree regression (DT) and two ensemble models are Random Forest (RF) and Gradient Boosting (GB) of base learner was the decision tree. The RF is a set of decision tree that is arranged parallelly to each other, the more the decision tree grows on every subset of the resampled version of original training data. Another ensemble model is GB, the base learner are decision tree which is arranged in series, every new decision tree has relative low error than the previous decision tree. We can understand the intuition of models in the model introduction. These ML models are implemented for two different datasets used for training, the first dataset (yellow) considered if core data is available as case 1, and for case 2, if core data is not available. The data set consider for case 1 is core data of well #1 is used as the target, and corresponding values of same well logs features are considered. A total of 96 data samples are used the workflow diagram is shown in figure 4.1.

For case 2 (figure 4.2) the training dataset of two well logs (well #2 and well #3) was used (blue). The target and features values are available for training for ML models which is provided by Statoil. Both cases are used for the prediction of K_v of well #1 (red). The best model was exposed to a new blind test dataset that is well #1.

4.2 Applied Model

Five ML models are implemented for the prediction of K_v from both training approaches after cleaning and processing the 7 conventional logs. We assume that two shallow learning models and three tree-type ensemble models can capture the pattern and hidden relationships between the seven input features and one output log (target). Further data-driven techniques, we researchers are also interested and applied empirical models for sandstone to investigate how far the empirical correlation function on our sandstone experimental/core data we can compare this in figure 6.5.

Moreover, a successful model should not only have high prediction accuracy but also good stability when exposed to a new dataset. Further, the best model evaluated to ensure the model stability for the robustness of modeling

(Case 1)

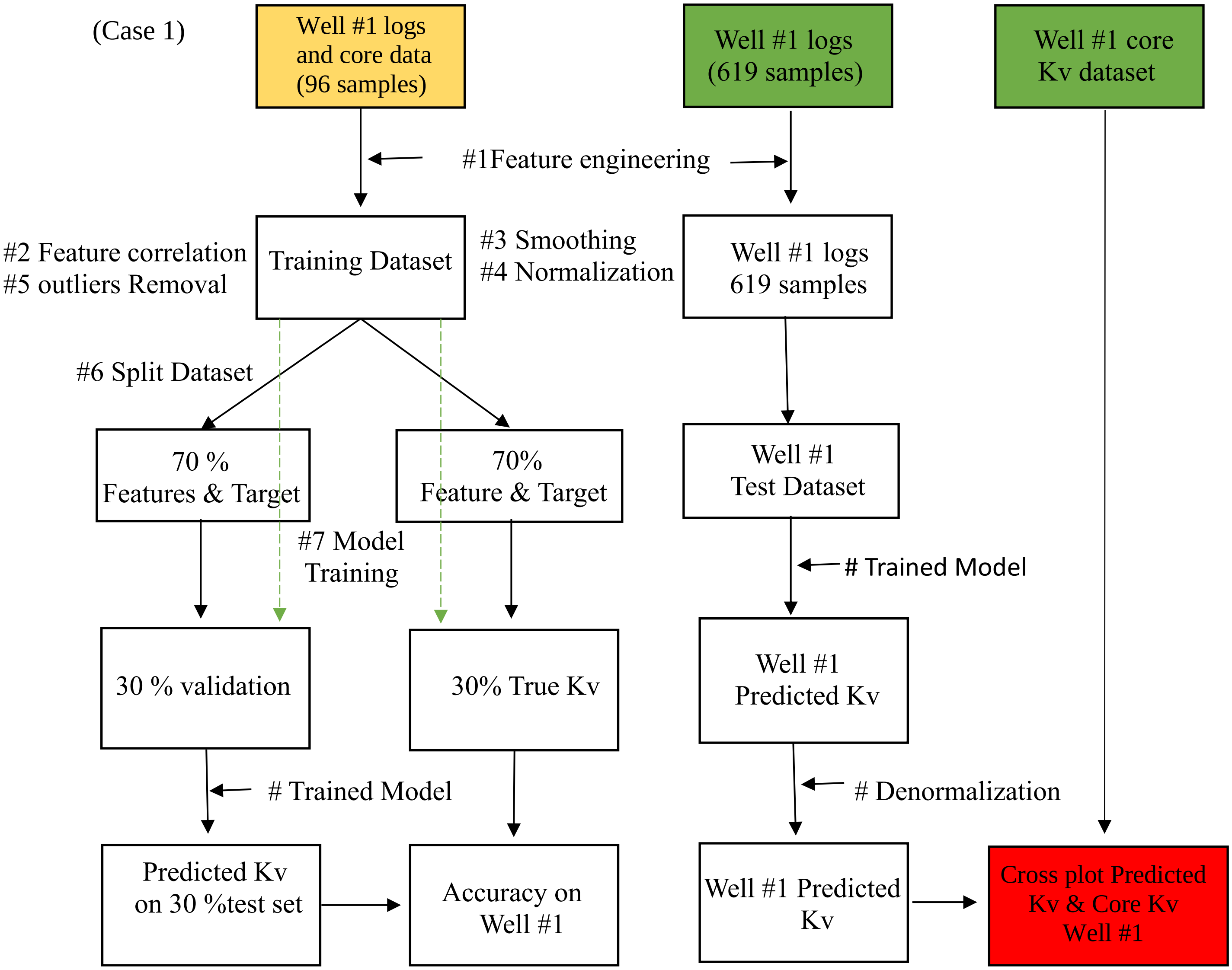


Figure 4. 1: The schematic diagram showing workflow of data processing and model training (case 1) from well log and core data, testing, prediction.

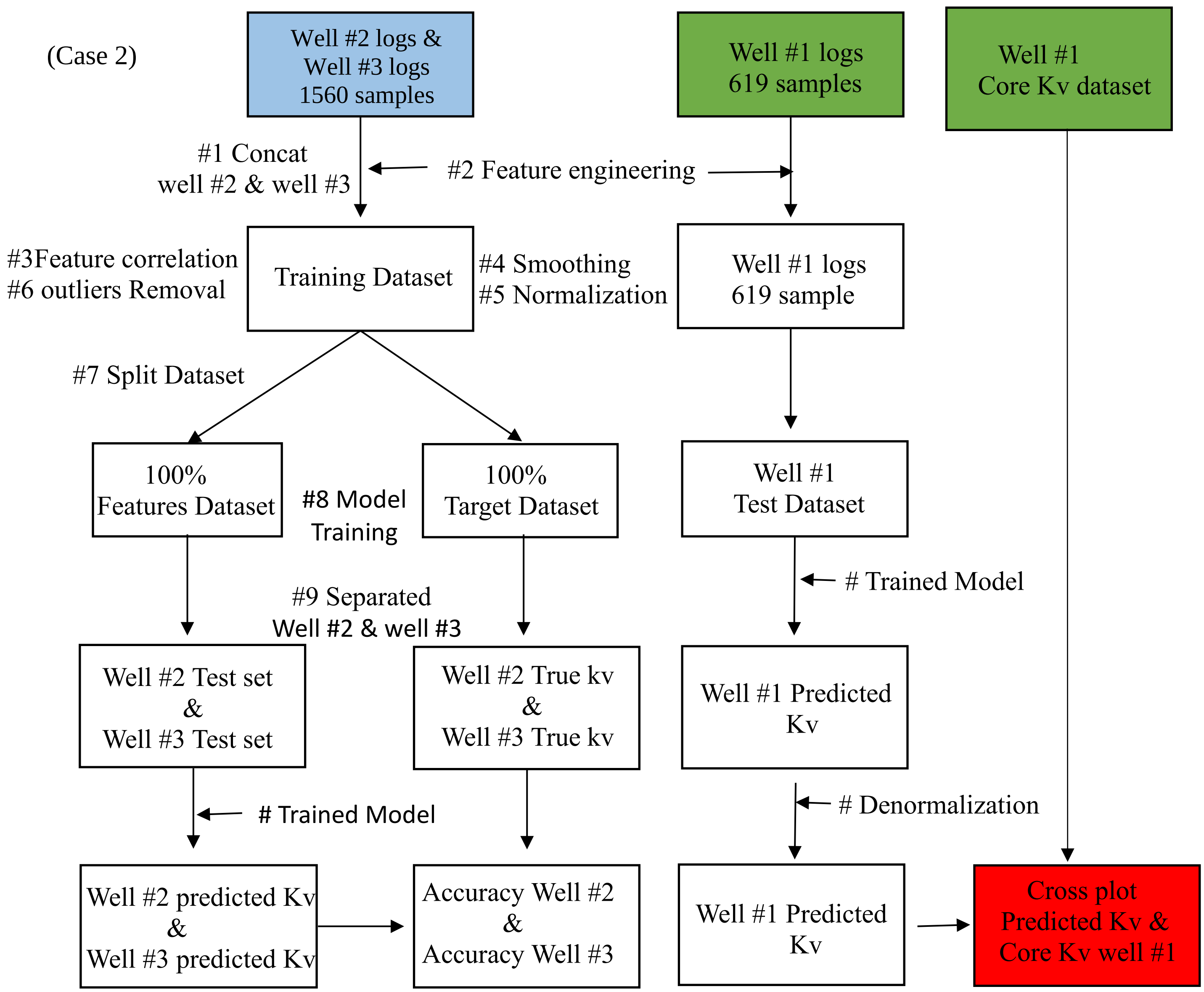


Figure 4. 2: The schematic diagram showing workflow of data processing and model training (case 2) from well logs, testing, prediction

4.3 The empirical model with core data

We consider the empirical correlation of K_v of sandstone formation which was driven by Shedid, (2019) which indicates the correlation of coefficient is 0.86. The model requires porosity and K_h, both parameters are directly proportional to K_v.

$$\mathrm{Kv} = 7.2 * \left(\sqrt{\frac{Kh}{PHIF}}\right)^{1.1094} \tag{4.1}$$

However, there is no universal correlation model are developed why because the petrophysical properties are highly dependent upon the depositional environment and diagenesis process of the sediments. Therefore the degree of permeability anisotropy is different for most of the reservoirs. In this model, we calculated the K_v values for each same from the given above correlation for Sandstone. The estimated K_v is a cross plot with core data samples of Hugin sandstone formation in figure 6.5.

4.4 Shallow models

4.4.1 Multi Linear Regression (MLR)

The linear regression is most useful when the features variable that is independent variable are y_1, y_2, y_3, y_4, y_5, and y_6, of the 6 input features are much influence on the target variable 'y' that dependent upon linear relationship. Multi linear regression trying to find a straight line or plan for seven dimensions that best fit for training dataset by calculating gradient/slop for each input features are given in bellow equation 4.3 are m_1, m_2, m_3, m_4, m_5, and $m_6$and manage to create a hypothesis in between dependent features and independent feature. Thus the same hypothesis was used for test data, mathematically we can state to minimize the residual sum of the squares between the observed targets in the dataset, and the targets predicted by the linear approximation simultaneously adjusting the bias/intercept ‘b’. Therefore the input/independent variable of our dataset is taken as features of well logs as mention above and the target/dependent variable is K_v.

The distance between the regression line and the data point represents the unexplained variation, which is also called the residual sum. The Residual sum is the squares between the observed and model-predicted target output values.

$$\text{Residual sum} = \sum_{0}^{i} (yi - \widehat{yi})^2 \tag{4.2}$$

Where, yi= data points of input feature, $\hat{y}$ is predicted values, y is target data (supervised learning)

$$y = m_1y_1 + m_2y_2 + m_3y_3 + m_4y_4 + m_5y_5 + m_6y_6 + b \tag{4.3}$$

4.4.2 Artificial Neural Network (ANN)

The neural network applications to subsurface characterization are elaborated by ANN, a commonly used machine learning model suitable for nonlinear regression (Jaiswal & Das, 2018). The architecture of the neural network system comprises an input layer, an output layer, and a few hidden layers that make up a neural network, each layer has unique input and output feed. The number of hidden layers and neurons in each hidden layer can be increased or decreased to change the neural network model's ability to fit data in higher dimensions. Each hidden layer has neurons with parameters (weights and biases) that perform matrix computations on signals computed in the previous layer and then an activation function. Each layer's activation function adds nonlinearity to the computations. We used the rectified linear unit (ReLU) activation function mostly looks and acts like a linear activation function. ReLU helps us to train deep networks as well as shallow networks with a multi-layered network with a nonlinear activation function using backpropagation. Backpropagation is the core of neural net training is back-propagation. It's a technique for fine-tuning the weights of a neural network using the error rate from the previous epoch (i.e., iteration). Overall our goal is to reduce error rates and make the model functional by fine-tuning the weights. Now we need to compile the ANN model to define the loss function/model errors as mean squared error and optimizer to converge the ANN model towards actual prediction.

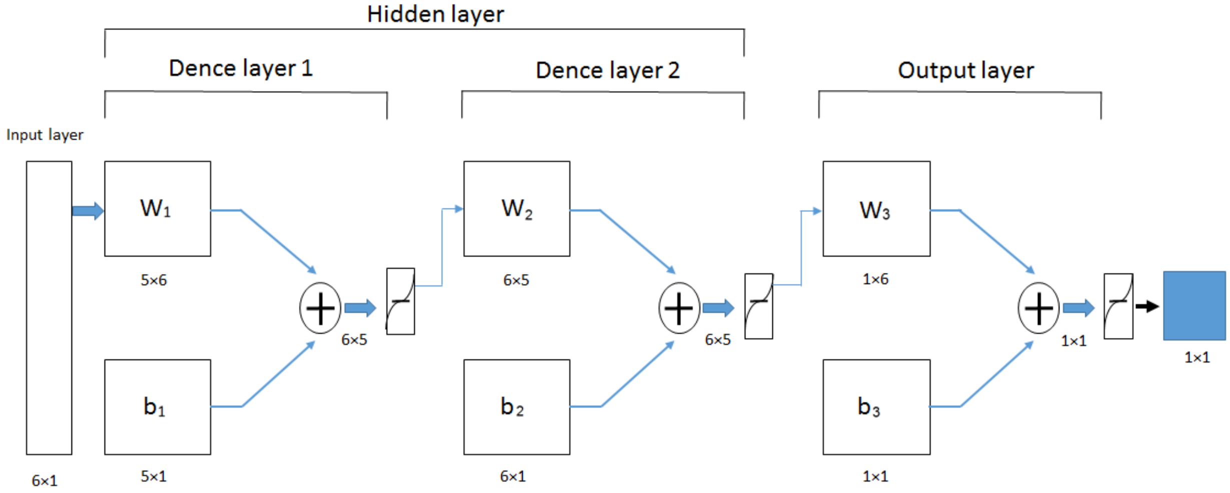


Figure 4. 3: Artificial neural network showing two hidden dence layers of 5 neural and 1 output layer.

We exhaustive grid search is used to measuring different learning speeds, number of hidden layers, number of neurons, and activation functions for the ANN model. An example of ANN architecture is shown in Figure 4.3. To be expected, there are six input logs and one output log (K_v). The input and output layers have 6 and 1 dimensions, respectively. The ANN model has two hidden layers, the first and second hidden layers have five neurons. The total number of parameters computed was 71 obtained from the model summary.

4.5 Tree-based models

One of the most prominent Machine Learning algorithms for predicting tabular and spatial datasets is tree-based. The tree-based model can be drawn like below. Starting from the top node, it divides into 2 branches at every depth level. The last end branches where they do not split anymore are the decisions, usually called the leaves. In every depth, conditions are questioning the feature values. The binary answer will decide which branch we are going to next. This process is repeated until we reach one of the leaves, which no longer splits. The prediction can be derived from the final leaf.

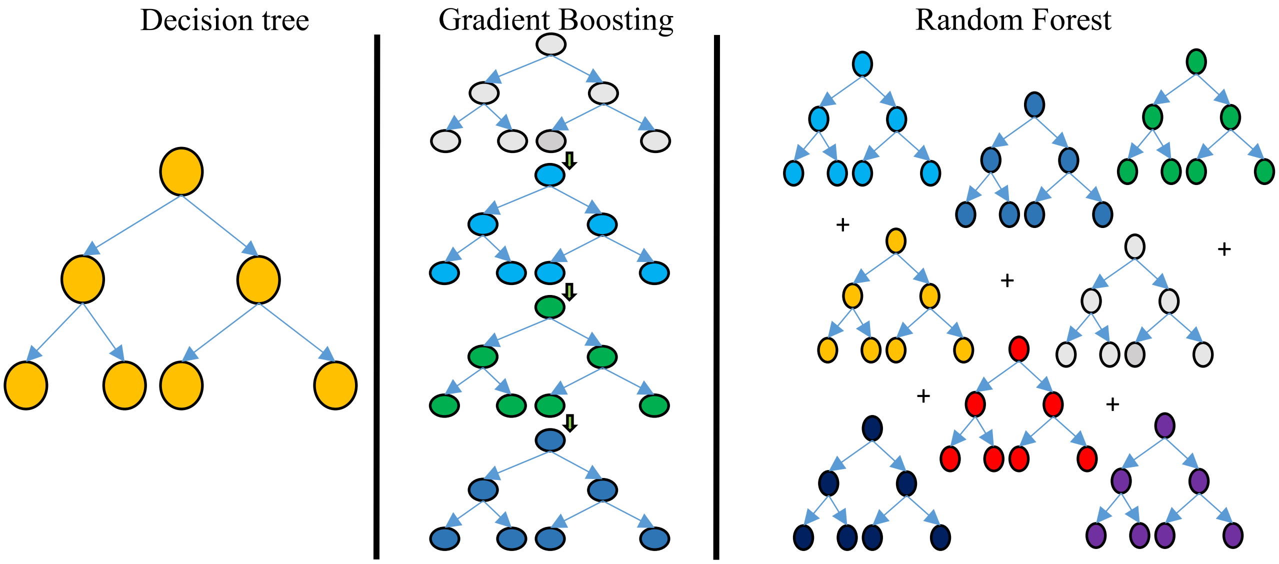


Figure 4. 4: Shows tree-based models are decision tree, Gradient Boostingand Random Forest.

4.5.1 Decision tree

A decision tree is the fundamental algorithm of ML a rule it breaks the training dataset into smaller and much smaller subsets in a recursive manner is called recursive partitioning. The construction is based on a hierarchical tree structure it is a non-parametric tree flow chart constructed as three layers/steps which are root node/the root, Nodes/internal nodes, leaf nodes/terminal leaf node. At the very beginning, the root node is started with asking questions/conditions on a set of input data at each leaf node the training data samples are further split into multiple internal nodes depending on the possible lowest value of mean squared error and finally, the leaf node shows a decision to made and an end node shows the outcome of a decision tree. Figure 4.5 (b) shows the original decision tree for better visualization of the tree we considering five subset samples of training data

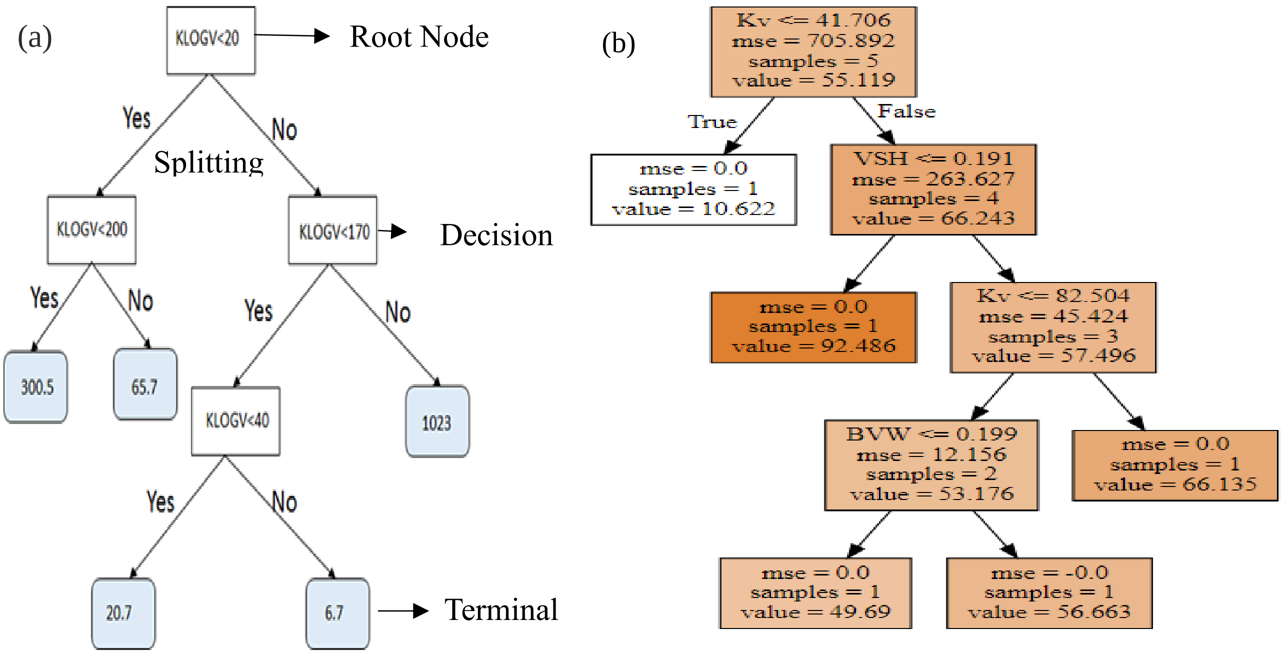


Figure 4. 6: (a) shows the ideal decision tree executing nodes by quries and binary split and (b) shows the original decision tree with five samples for training.

sample which is executing nodes by answering the query, for every query there are answers along with accuracy mean square error.

However, for real-world problems, no leaf node can give a hundred percent accuracy based on a decision if overfitting is avoided. The decision tree computes impurities based on total true and false, we come up with Gini impurities state that the total minus squared probability of "yes" minus squared probability "no", However, if the samples of yes and no are not equal then we come up with a weighted average. Hence the lowest Gini impurities can predict a better much better result.

4.5.2 Random Forest

A Randon forest is an ensemble technique of supervised machine learning and its non-linear nature capable of capturing the hidden relation/pattern between the features and target. One decision tree may have high variance, we can simply reduce the high variance when we combine all the single decision trees to form a group of trees. The resultant variance may have low as each decision tree is trained on a subset of training data. The different subset of training data is created from the bootstrap re-sampling technique, each subset of training data is used to train different single decision tree. The final output is the mean of output/vertical permeability from every single decision tree is called aggregation. In the given below figure 8, we can see the random forest model consists 'n' number of decision trees formed and trained by different subsets of original training data. Further, we can visualize one tree of the random forest by considering five data samples for training the random forest model.

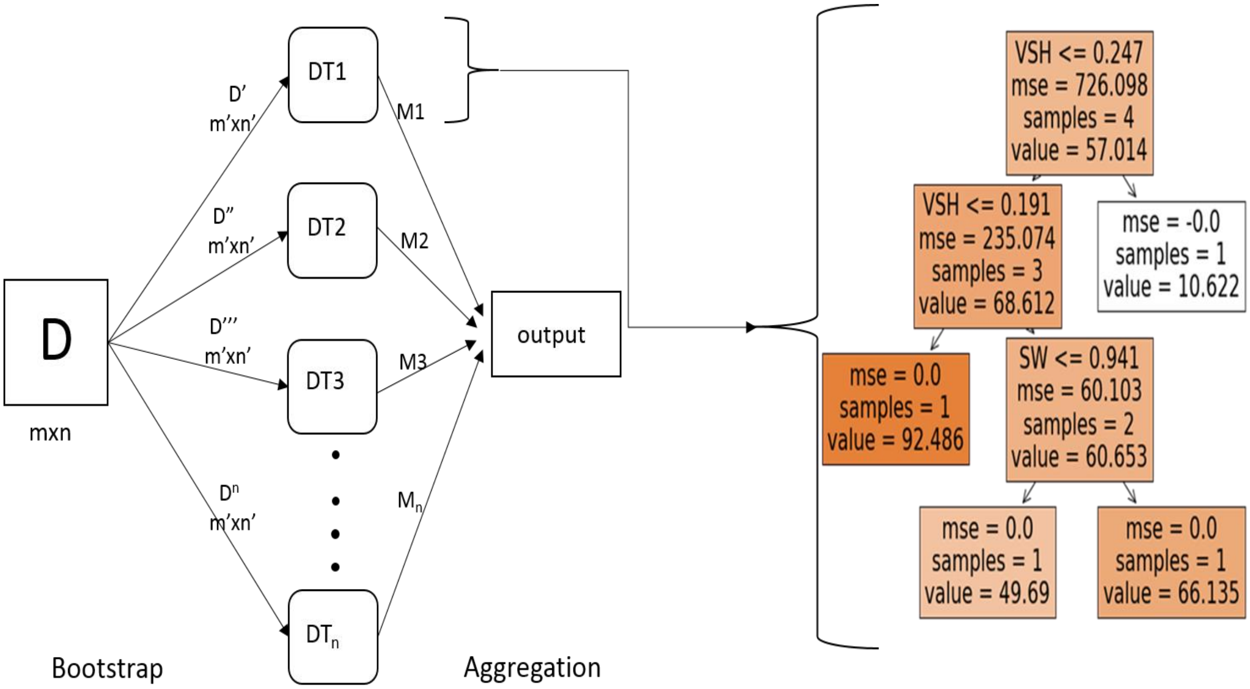


Figure 4. 6: The random forest contains 'n' number of a single decision and visualization of a single decision tree of a random forest.

The simple basic idea was to parallel combine multiple decision trees as base learning models. Bootstrap is the process of randomly picking rows and features from a training dataset to create a subset of training datasets for each model and aggregation helps to combine the output result by considering the mean all decision tree model is called bagging.

4.5.3 Gradient Boosting

Gradient boosting techniques (GB) are powerful and popular machine learning algorithms which is a sequential ensemble approach of combining base learning to multiple shallow trees in sequence with each tree learning and improving on the previous one. Although the shallow trees among all are rather weak predictive models (typically decision trees). The predictions from the individual base models that make up the ensemble are combined to make new predictions (e.g., by averaging in regression). The key concept behind boosting add new models to the ensemble sequentially. Boosting tends to globally minimize the residual error by accounting for the tradeoff of the bias and variance from starting with the base learning model to the sequence of shallow weak tree models. Shallow tree means relatively few splits (1-6 split) and having few layers tree is represent as week learner. Understand the important structure/components of boosting.

The base learners: The base model can use decision trees and random guess near values to the dependent variable (K_v). However, the boosting framework iteratively improves the dependent feature by reducing the pseudo residual errors from every iteration, so base learners have maximum pseudo residual error. Training week model one whose error rate is only slightly better than random guessing. The idea behind boosting is that each model in the sequence slightly improves upon the performance of the previous pseudo residual error. By focusing on the rows of the training data where the previous tree had the largest errors/residuals. Sequential (series) based approach for training concerning evaluating errors**,** boosted trees are grown sequentially each tree is grown using information from previously grown trees to improve performance. By fitting each tree in the sequence to the previous tree’s residuals, we’re allowing each new tree in the sequence to focus on the previous tree’s mistakes, for every new tree there is reduce in an error of GBR.

1. Fit a decision tree to the training data: $(x)\ f1=y$
2. Then fit the next decision tree to the pseudo residuals of the previous: $(x)h1= y - (x)f1$
3. Add this new tree to our algorithm: $(x)\ f2 = (x)\ f1+(x)h1$
4. Fit the next decision tree to the residuals of $f2$: $(x)h2=y-(x)f2$
5. Add this new tree to our algorithm: $(x)f3=(x)f2+(x)h2$

The final model here is a stage-wise additive model of *k* individual trees:

$$f(x) = \sum_{k}^{K} (x)\, f^{k} \tag{4.4}$$

Where *x* and *y* denote input/features and target (K_v) respectively, *f1…..fn* is the capture pattern between inputs and target of the model *(1…..n), h1* is the computed error of model 1, and so on.

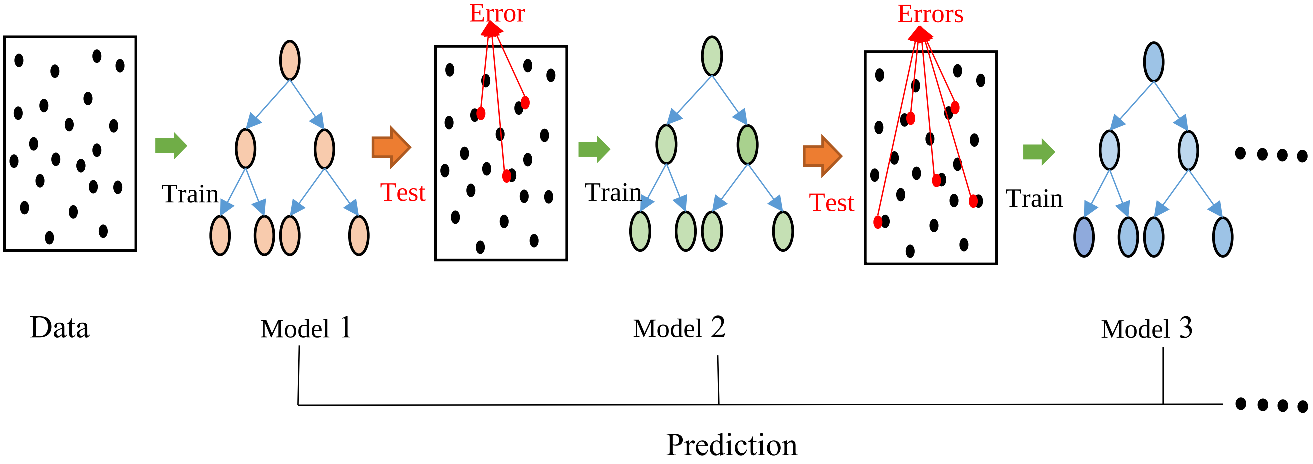


Figure 4. 7: Gradient boosting with multiple trees, adding tree by reducing mean squared error.

Continue this process until the model converges to global minimum loss and reduces the residual errors, however, we need to take care of the overfitting problem and maintain low bias and variance, and therefore the Gradient descent algorithm plays a crucial role to minimize the function of the residual errors.

CHAPTER- 5: PERMEABILITY PREDICTION WELL LOGS DATA AND CORE DATA

5.1 Case 1: Model training with well logs

After data pre-processing the training data distribution is normal and finally, we train the models in consideration of depth dependency. The training data is concatenated of well #2 and well #3, the left data sample for training is 1405 concerning depth and well number. However, we trained and fit the models with inputs to the target. we used both well #2 and well #3 for model validation with different model evaluation metrics. The cross-validation technique is applied while model training to ensure the model behavior.

Train – test split, ensure the data was chosen with the same mnemonics and unit for all training and testing logs data. For model testing and development, the two different well logs data are used after processing we left with 1405 data samples. Trained models with 100% training data in considerations of depth dependencies, it can help effective training of the models, generalized and capture the hidden relations between features/input and target. For validation of the model again in considerations of depth dependencies we separated both wells of training according to the well name. Therefore, we tested the individual model for both wells (well #2 and well #3).

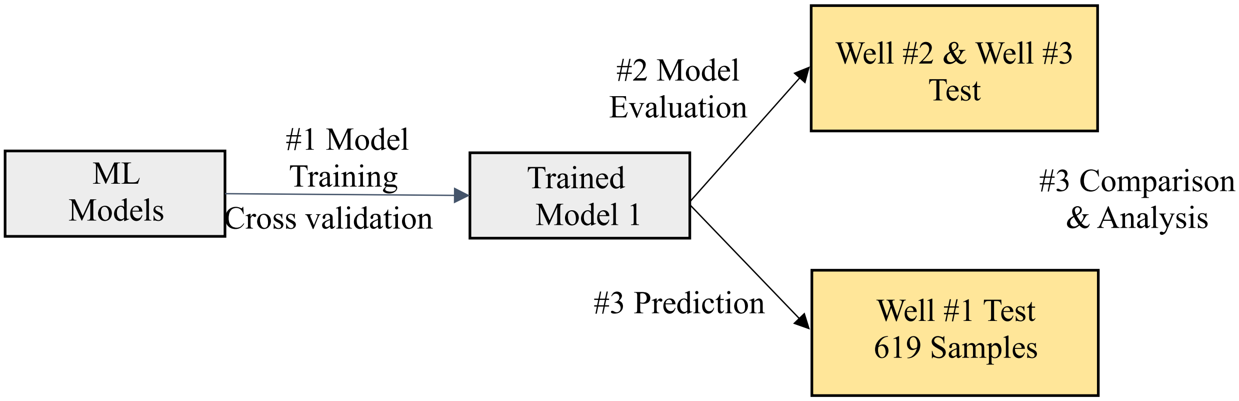


Figure 5. 1: Machine learning training (case 1), cross validation, testing on well #2 and well #3 and well #1 tested on core data.

5.2 Case 2: Model training with core data corresponding to well logs

The core dataset of well #1 is available for model training, the core data available from the depth range of 3837 meters to 3965 meters having 96 data samples of $K_{v.}$ The corresponding well logs data of the same well used for model training. However, for the model training, we consider a target of core K_v. Therefore, the total number of 96 data samples available for model training. However, the availability of small training data can influence the ML model (Zhang & Ling, 2018). We can analyze the fundamental of the predictive capability of all five ML models and finding the best model among all the models. For this model, we had implemented all the data preprocessing methods except train test split for model training. A total of ten outliers were detected and removed from the training dataset, however, we left with only 86 training data samples.

Train – test split, for this model was done for model development and testing. The 86 training samples are split into 70 percent model training (60 data samples) and 30 percent for the testing dataset (26 data samples).

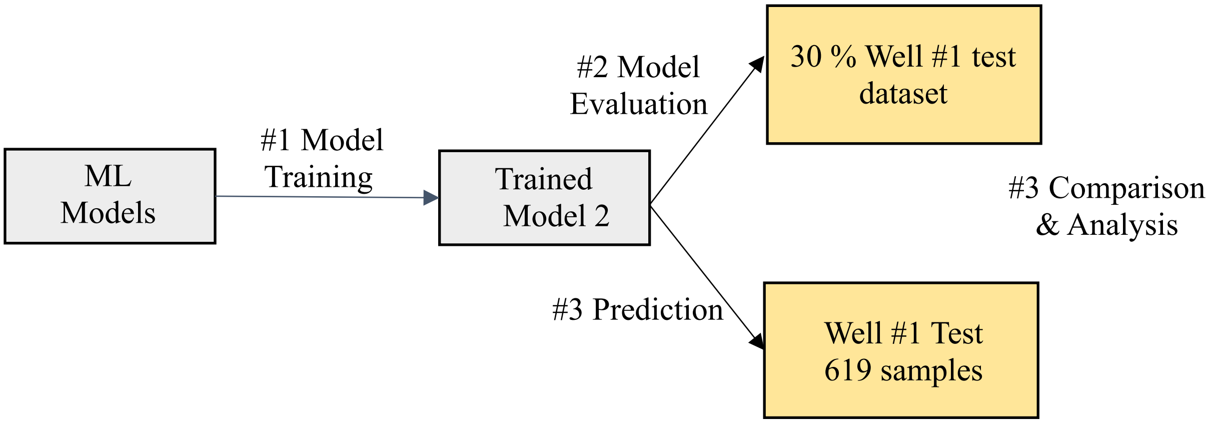


Figure 5. 2: Machine learning training (case 1), testing on well #1 logs and core data for model evaluation and final prediction for well #1 test data contain 619 samples.

5.3 Metric used for model evaluation

Model evaluation is the basic building block of the ML model which states many things to us, this is useful for the model development process and helps us find the useful model that represents our data and how the model will work well in the feature. To know the best performing model and their behavior, therefore, it is necessary to use more than one model evaluation method. However, we used four different metrics for model evaluation. For a better understanding of ML models, we used model evaluation on test data. These four metrics are used R^2 (coefficient of determination), MSE (mean squared error), MAE (mean absolute error). R^2 coefficient states that the ability to describe the individual regression models calculated from the sum of the squared error to the sum of squared total or simply measure the variance obtained from the model, mathematically express as-

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left(Y_i - \hat{Y}_i\right)^2}{\sum_{i=1}^{N}(Y_i - \bar{Y})^2} \tag{5.1}$$

Where Y and $\hat{Y}$ both are the original value and predicted value of KLOGV and the remaining $\bar{Y}$ is the mean original value. The mean squared error (MSE) can state that how close the data points to the best regression line fit and measures the residual error (Euclidean distance between data points to the regression) and squaring them this is mathematically express as-

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \hat{Y}_i\right)^2 \tag{5.3}$$

Mean absolute errors (MAE) measures the average magnitude/absolute of the residual error this metric is similar to RMSE as calculating the residual error, but there is a difference since the residual errors are squared before averaged, Therefore the RMSE measures a relatively high weight to large error thus it's more useful when the large residual error is expected. MAE is mathematically express as-

$$\mathrm{MAE} = \frac{\sum_{i=1}^{N}\left|Y_i - \hat{Y}_i\right|}{N} \tag{5.4}$$

The idea is behind calculating different types of residual error to investigate the loss whether is very minor as we can see in the below section.

5.4 Model stability investigation

There are enough data samples to perform cross-validation techniques for trained ML models from well logs to obtain the model's performance on every fold of data. Cross-validation can help us to understand and generalize the ML models where overfitting (overlearning) and underfitting (not able to capture the underlying trend) to overcome this problem or avoid an increase in errors and bais so that the model learning is effective. Further, we implement the 5-fold cross-validation shown in figure 5.3 to investigate the stability of the machine learning model. 5-fold cross-validation technique split 100 % of the training data into five equal parts. The model is trained 5 times, every time the model is trained with 4-fold that is 80% of the training data, and validate with the remaining fold of data that is 20% of total data. The cross-validation techniques is used for the best model obtained from the metric score to investigate stability at every 20 % percent of training data to ensure about model is trained to generalize over the entire available dataset. This is very important to generalize the entire training data, otherwise, the model may lead to not stable for test on the blind test dataset. For this cross-validation for every fold builds a model on every fold that is five models with as discussed above the ratio of training and testing data tested on the same corresponding model. The mean squared error is measured for every five models to obtained overall accuracy will be average of all models.

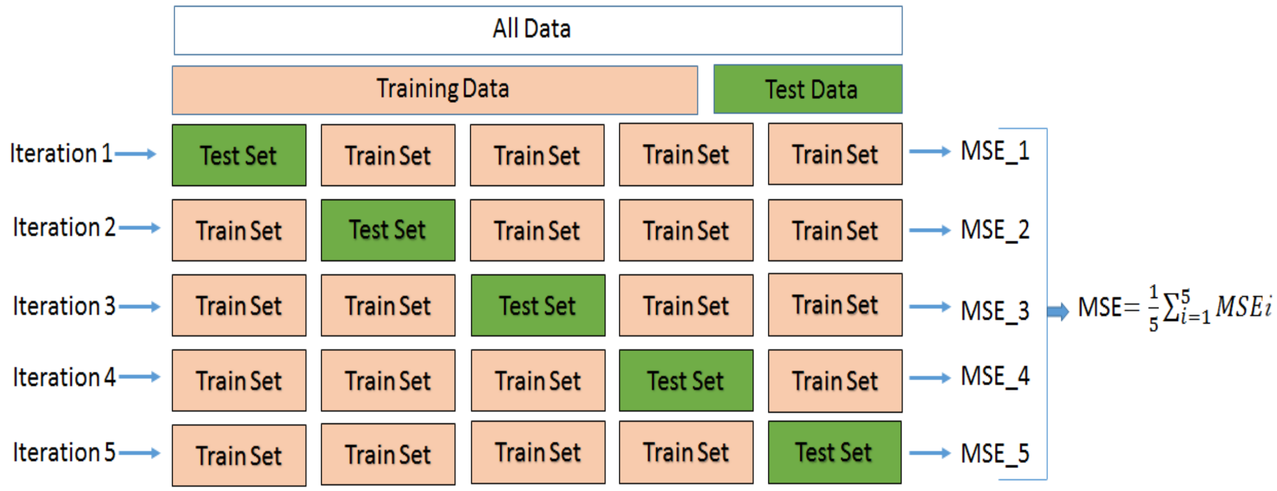


Figure 5. 3: Shown training data divided among in 5-fold, each fold split 80 % training data and 20 % test data.

This analysis helps us to understand the model stability whenever the machine learning model is exposed to a new dataset as or blind data the model works effectively. However, if the model learns false or not able to capture the hidden relation then the mean squared error (MSE) for each fold or model are varies drastically therefore the mean of all model MSE is not viable. The above experiment shows the stability of the model.

CHAPTER-6: RESULTS AND DISCUSSION

6.1 Case 1: Models trained with well logs

The performance of all five models is compared with all four metrics as introduced in the above section. These five models are tested for well #2 on 4040 meters to 4180 meters depth interval with 904 samples and simultaneously in same way tested for well #3 on 3700 meters to 3800 meters depth interval with 656 samples, there for the total samples for training the models are 1560 sample points including outliers. After the data preprocessing section, we left with 1405 sample points by treating the training data with one class Isolation forest as discussed in the above section. This time our goal is to find the best ML model by investigating all performance metrics and deployed the same ML model for blind testing data as well #1 on 3825 meters to 3920 meters depth interval with 619 samples. At the time of model training with 5-fold cross-validation to avoid over learning to ensure the robustness of the models as well stability.

The performance of the prediction evaluation is based on three metrics which are shown below listed in the following tables 6.1. The columns of the table are different model implemented and the row of the same table represent three different metrics used to evaluate every model. The first row and first column represent the R^2 metric used to evaluate the MLR model gives accuracy on the test dataset of well #2 and well #3 are 0.96 and 0.93 respectively. The R2 of shallow ANN for well #2 is 0.96 shows the goodness in the result but failed to give high accuracy on well #3 it gives only R2 value is 0.89. Simultaneously the residual error for well #2 shows high values and validates the above statement however shallow ANN fails for this data-driven model, for the shallow model the MLR performs well. The tree-based model performs super, the ensemble models GB and RF show the same very accurate R2 value for well #2 is 0.9, and 0.98 respectively.

Table 6. 1: Prediction performance of the 5 ML models using 3 metrics. The table shows the performance of all models on the testing dataset from well #2 and well #3.

Metric	Test set	MLR	ANN	DT	GB	RF
R2	Well 2	0.96	0.96	0.97	0.98	0.99
	Well 3	0.93	0.89	0.93	0.94	0.94
MSE	Well 2	0.01	0.007	0.013	0.01	0.007
	Well 3	0.05	0.04	0.05	0.05	0.04
MAE	Well 2	0.02	0.03	0.02	0.05	0.029
	Well 3	0.04	0.03	0.04	0.06	0.03

However, the accuracy on well #3 of both ensemble models are the same R^2 is 0.94, from the correlation coefficient the RF model shows high accuracy, however, the other metrics MSE and MAE in figure 6.1 and 6.2 showing more error for GB as compared to RF and validate the above statement of R^2. The other tree-based model DT observed R^2 values on well #3 is 0.97 is less accurate than the ensemble model on the DT model as compared to RF. Therefore, for above discussion from the interpretation of figure 6.1 and 6.2 validate the RF models is more accurate than other models.

Therefore, for this data-driven modeling for K_v log over depth the RF model leads to gives the highest accuracy. However, we must have to ensure the stability of the RF model is preferred. In the below section we introduce the cross-validation technique to investigate the stability of the RF model is necessary.

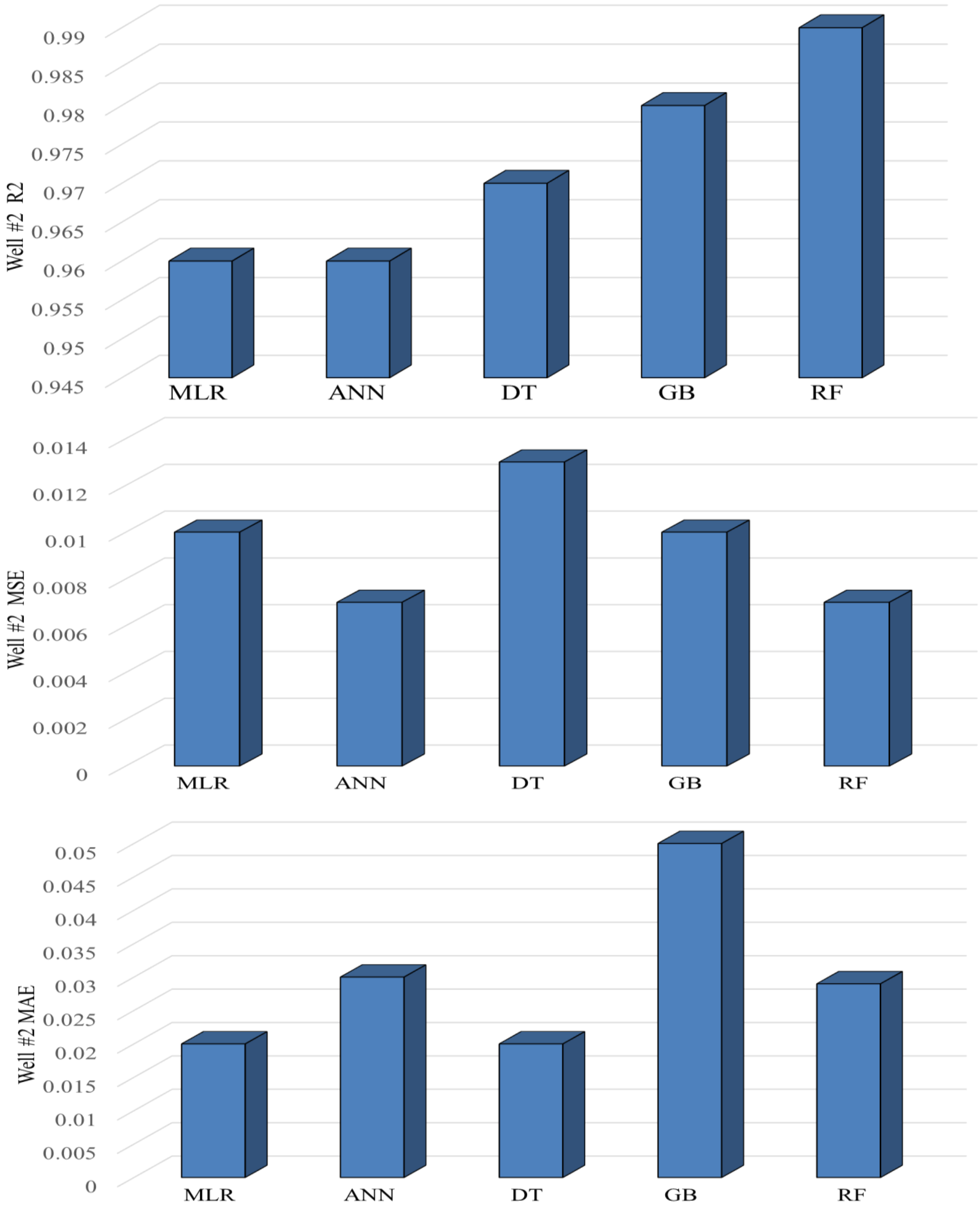


Figure 6. 1: The metrics were evaluated for all models. The figure shows the model performance metric on test data from well #2.

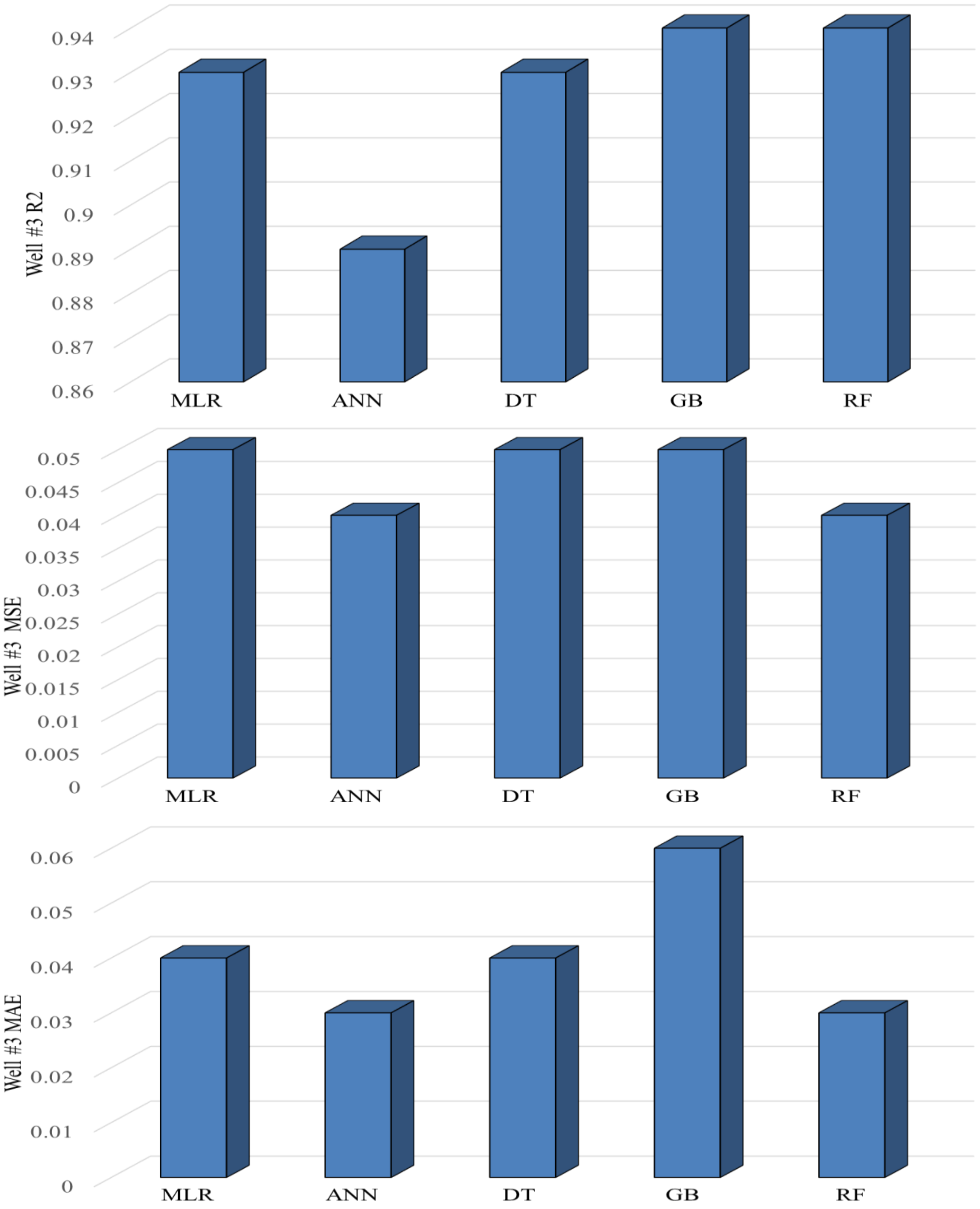


Figure 6. 2: The metrics were evaluated for all models. The figure shows the model performance on test data from well #3.

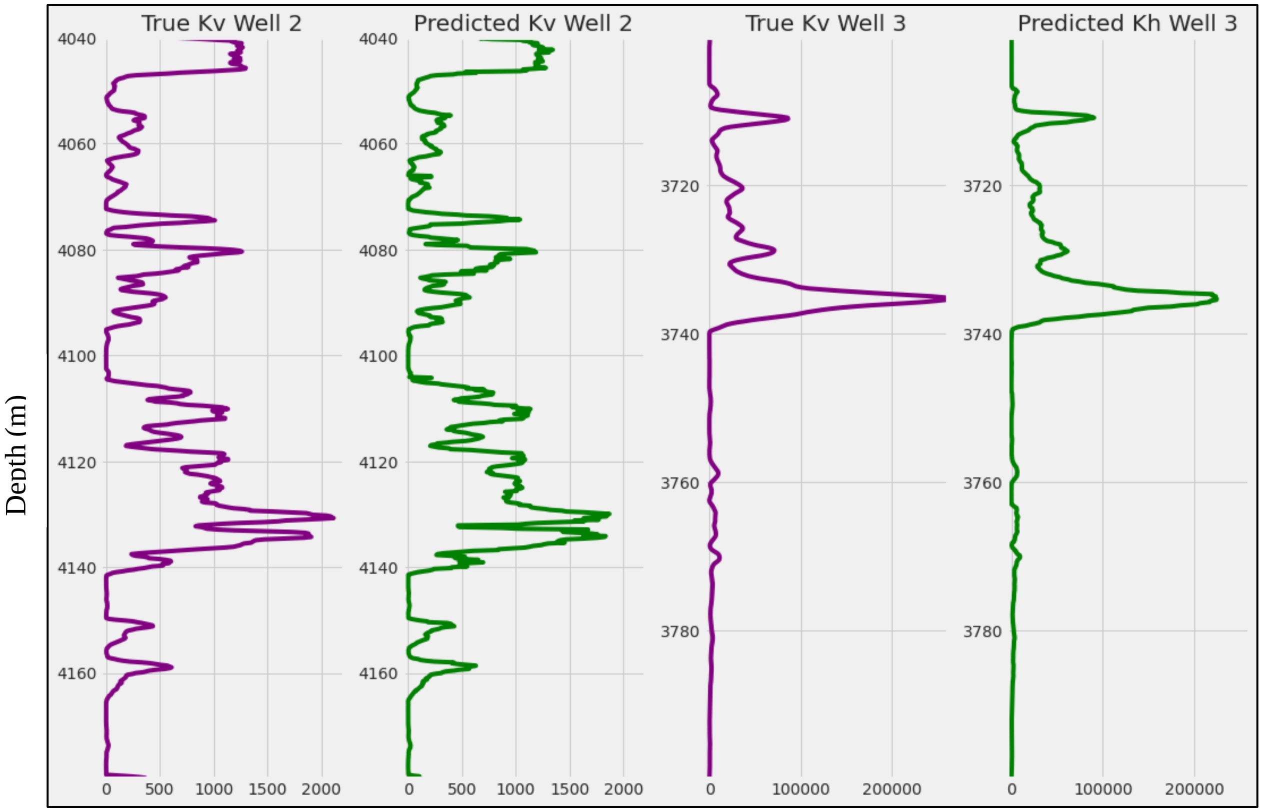


Figure 6. 3: Comparison of true and predicted Kv, first two tracks for well #2 and last two-track for well #3, after deployment of RF model on both well respectively.

The actual and predicted K_v logs are plotted in figure 6.3 to intuitively understand the model deployment with depth dependencies and understanding of the performance of the RF model. The first two tracks are true and predicted values of K_v and the prediction on the test data with high accuracy shows evidence in figure 6.3 predicted K_v. However, the robustness of the trained model can be observed second track, prediction on well #3 test data is good enough this shows the model ready to test and predict on blind test data.

6.2 Case 2: Models trained with well logs and core data

Another training data set was considered for all five models, after outliers removal, the well logs and core training data set had only 86 samples for models training. The small amount of data influences all machine learning models, especially the ANN not stable same, therefore ANN model got very low accuracy in comparison of other models. The reason for low accuracy as compared to case 2 for all models in table 6.2 may be very low-resolution training data only 86 data samples from the depth range of (3837 m to 3963 m). influences the stability of ML models.

Table 6. 2: Prediction performance of model evaluated using metrics. The table shows the performance of all models on the 30% testing dataset from well #1

Metric	**Test set**	**MLR**	**ANN**	**DT**	**GB**	**RF**
R2	well #1	0.73	0.21	0.69	0.83	0.87
MSE	well #1	0.24	0.73	0.28	0.15	0.12
MAE	well #1	0.36	0.69	0.4	0.3	0.24

The R^2 of the shallow ANN model shows very low accuracy on test data. In the above table 6.2, the maximum accuracy was obtained from Random forest (RF) ML algorithms. Moreover, for well logs and corresponding core K_v training data, the best model obtained from the metric evaluation are RF. However for cross-validation technique not has been used due to low training samples. Similarly, from case 2, the predicted K_v is cross plot with core data shown in figure 6.5.

Similarly, the tree-based modes are performed well, and the RF is the best model that performs best on considered training data. Moreover, it shows low MSE and MAE, hence the RF model is used for the prediction of Kv, and further predicted log was obtained from pass well #1 test data (619 data samples) and cross plot against core data shown in figure 6.5.

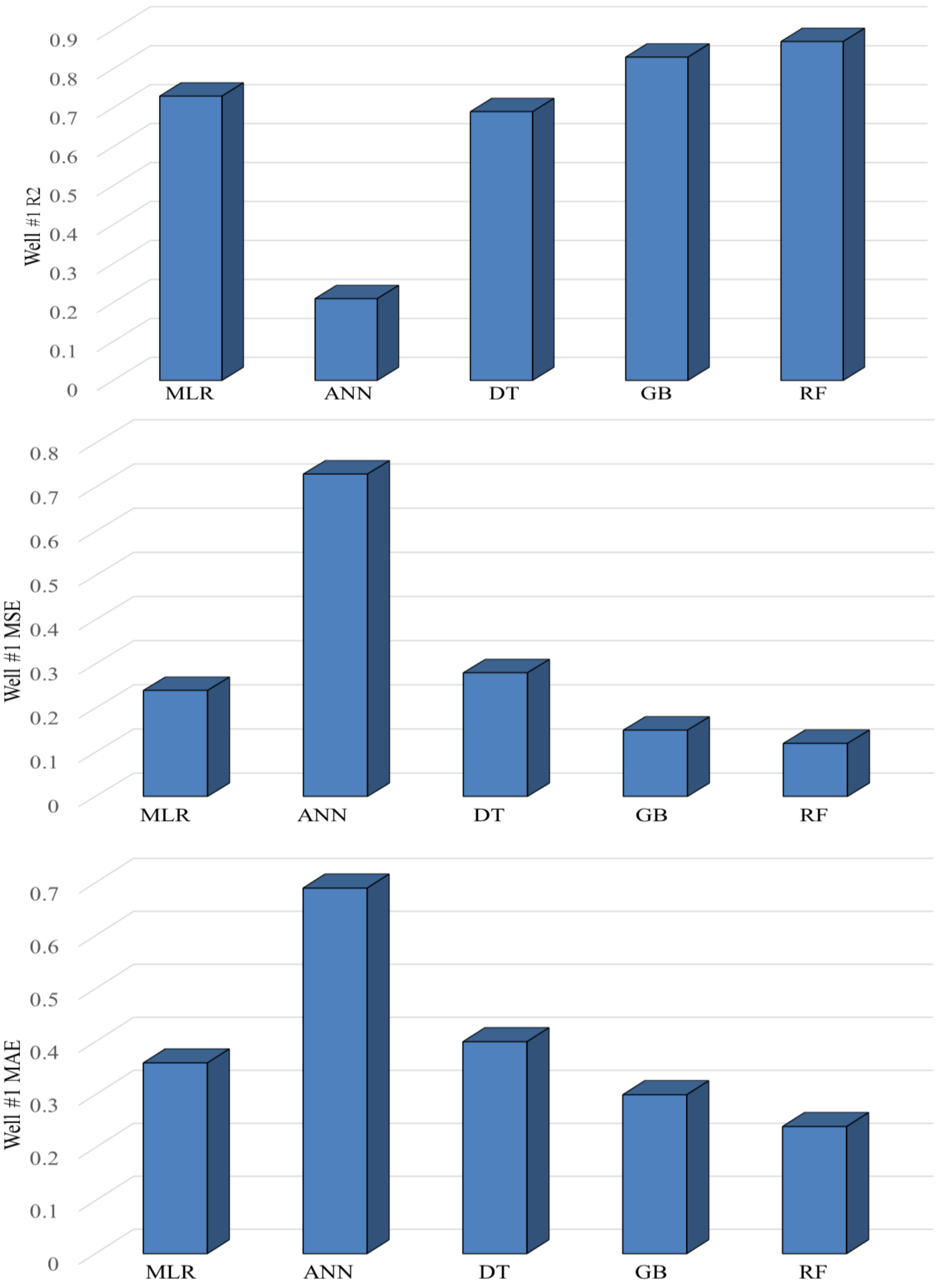


Figure 6. 4: The metrics were evaluated for all models. The figure shows the model performance metric on test data from well #1.

6.3 Cross plot analysis

In the above section, the RF model represents the best ML model. To further validate the predicted K_v with core data thus evaluation of RF model's prediction, compared our model best RF with core data (experimental data) along with the same depth interval. As above introduces the reader to the core data description in the above section. The plot of core K_v data sample points along with the predicted K_v from obtained best model. Visualize the plot in figure 6.5 first track shows the cross plot of empirical model K_v (red plot) and core $K_{v.}$ The empirical model K_v flows the trend of core K_v and satisfies the empirical correlation of sandstone. At the depth of 3850 meters, the empirical K_v is not able to compute the maximum values of $K_{v.}$ Therefore need to develop a new empirical correlation for the sandstone of Hugin formation. The second track represents a cross plot of core K_v and predicted K_v from well logs, and flows the trend of core data including the prediction from the best RF model which is trained from two well logs. The trend of the predicted blue line strictly follows the black dot as core data points. The RF model is expected to predict maximum values of K_v, as well as some of the minimum values, which are also intersected as compared to the empirical model. This represents the robustness of our RF model for blind test data (well #1). The interpretation of only K_v can give four different formation zone with high variation of K_v. The zonation of reservoir for pay zone interest is another aspect of the research.

The green Kv log (track 4) is predicted from the best model which is trained using well #1 and core data Kv cross plotted against smoothed core Kv with an average moving filter (track 3). The model is not as perfect as the model trained with good numbers of training samples. Therefore, the derivation of empirical correlation for Hugin sandstone formation required high-resolution Kv prediction to develop a stable correlation. The predicted Kv from the trained model with two well logs is considered for empirical correlation estimation.

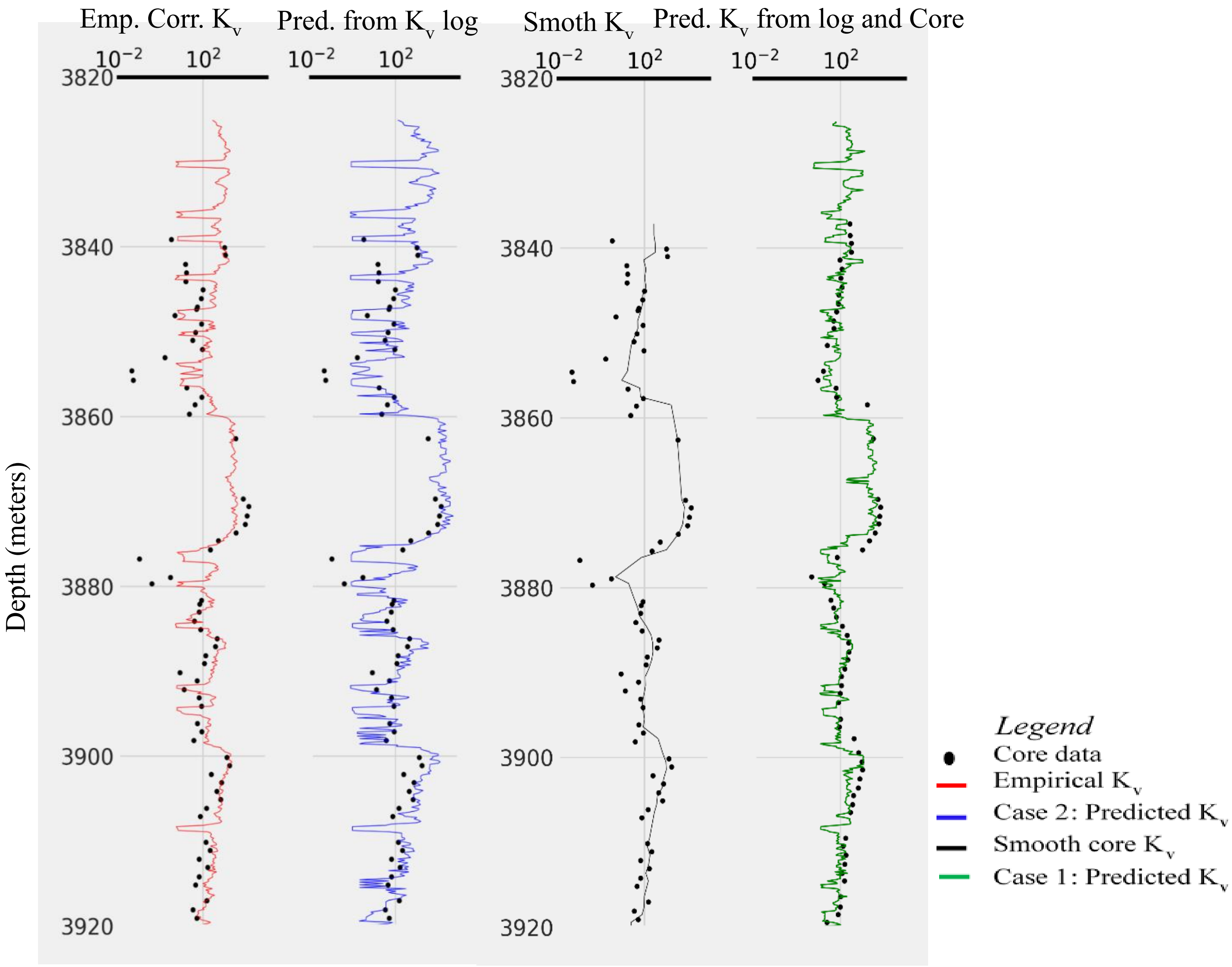


Figure 6. 5: Cross plot of core Kv with the empirical model Kv and predicted.

6.4 New empirical correlation

A new empirical correlation for Hugin sandstone formation can help to estimate the K_v from porosity. From this drive, new empirical correlation can save computational time and cost to avoid all the number of implemented techniques which is used above for prediction for K_v. Therefore the new empirical correlation equation 6.2 can help to estimate the K_v in a very simplified way. As we know the relation between permeability and porosity is direct and linear. To develop empirical formula such as linear correlation can be a suitable method by using linear polynomial fitting. We used the ordinary least squares (OLS) model in which the dependent variable is the predicted K_v and the independent variable is well #1 PHIF. The objective of the OLS model is to reduce the residual loss of the predicted K_v as real values to the estimated values from the regression line as we can see in figure 15. The model summary describes the correlation coefficient as R^2 is 0.414 and the constant value is -1.2816. The empirical correlation equation we found is shown in equation 11. Shedid (2019) was computed K_v correlation for sandstone formation using core data and found the correlation coefficient was 0.493 (Shedid, 2019).

$$\mathrm{Log}_{10} Kv = 18.698 * PHIF - 1.281 \tag{6.1}$$

$$Kv = 10^{(18.698*PHIF-1.2816)} \tag{6.2}$$

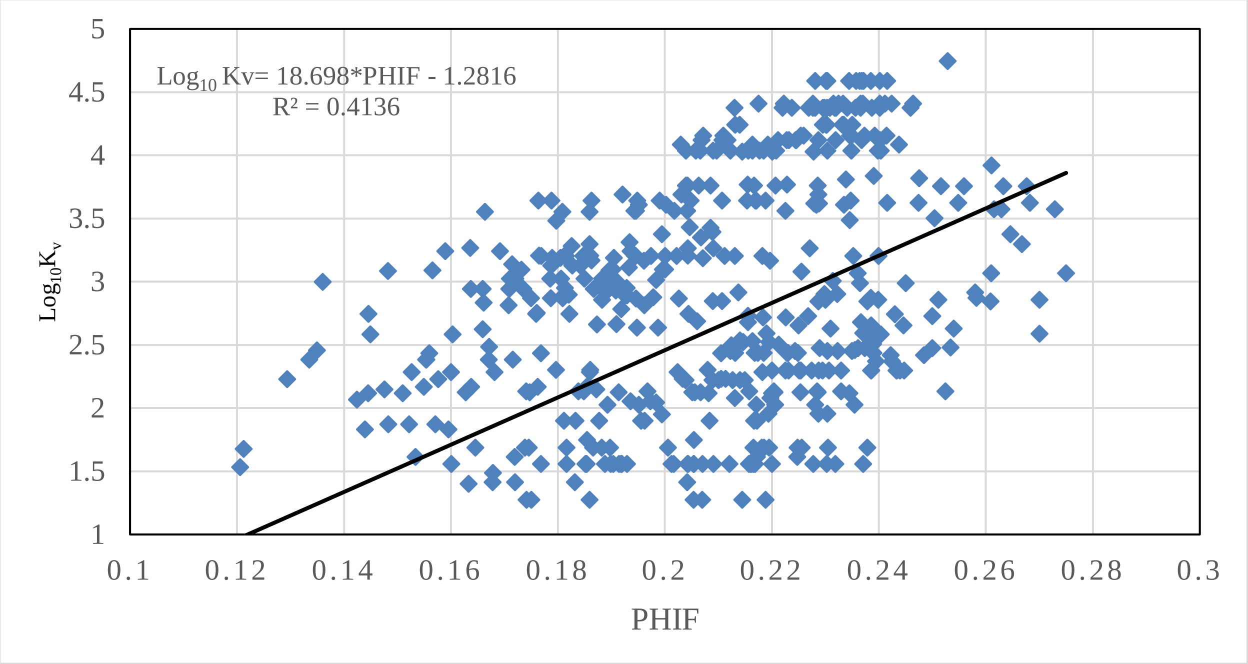


Figure 6. 6: The plot of Kv vs PHIF and the developed new correlation coefficient is 0.4136.

The deposition condition and variety of environment are not the same for most of the reservoirs, so most of the reservoirs are may not homogeneous, only varying degrees of heterogeneity within a short distance. The fluid displacement behavior during secondary recovery or enhance oil recovery the reservoir engineer must understand the both vertically and laterally variation of reservoir permeability. The study of K_v/K_h represented the anisotropic ratio of a reservoir (vertical variation of permeability). Thus, the graphical representation in the below figures can help us to understand the anisotropic of well #1.

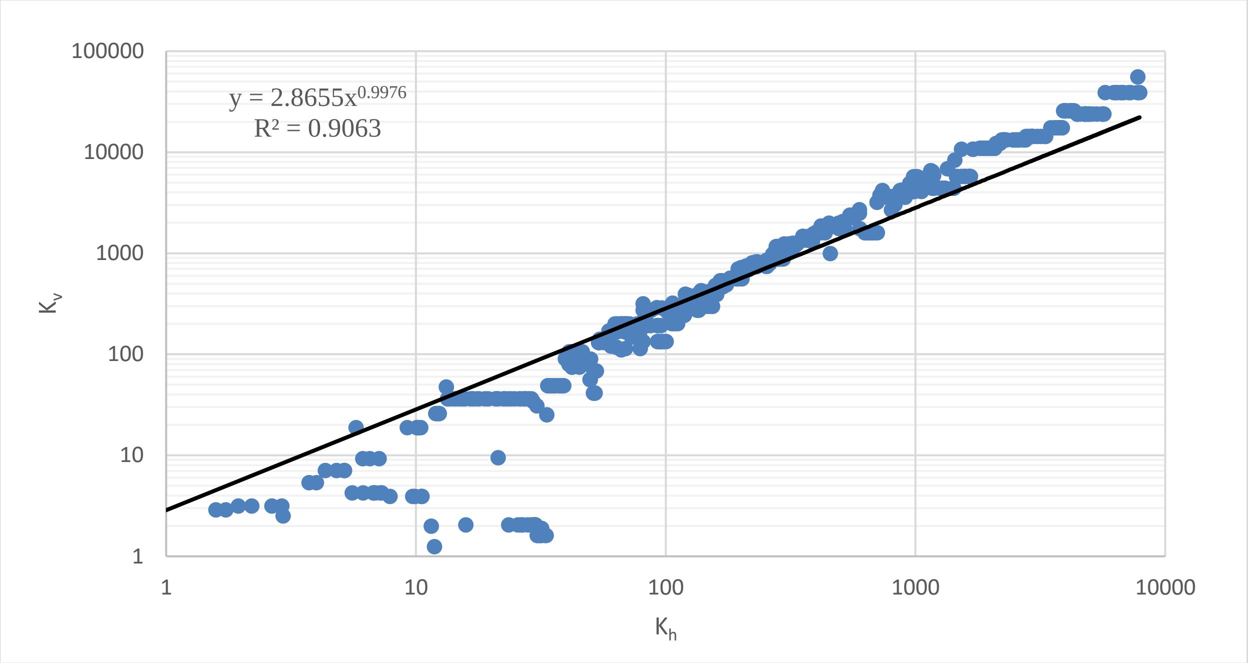


Figure 6. 7: Developed a new correlation equation for vertical permeability and hydraulic mean radius

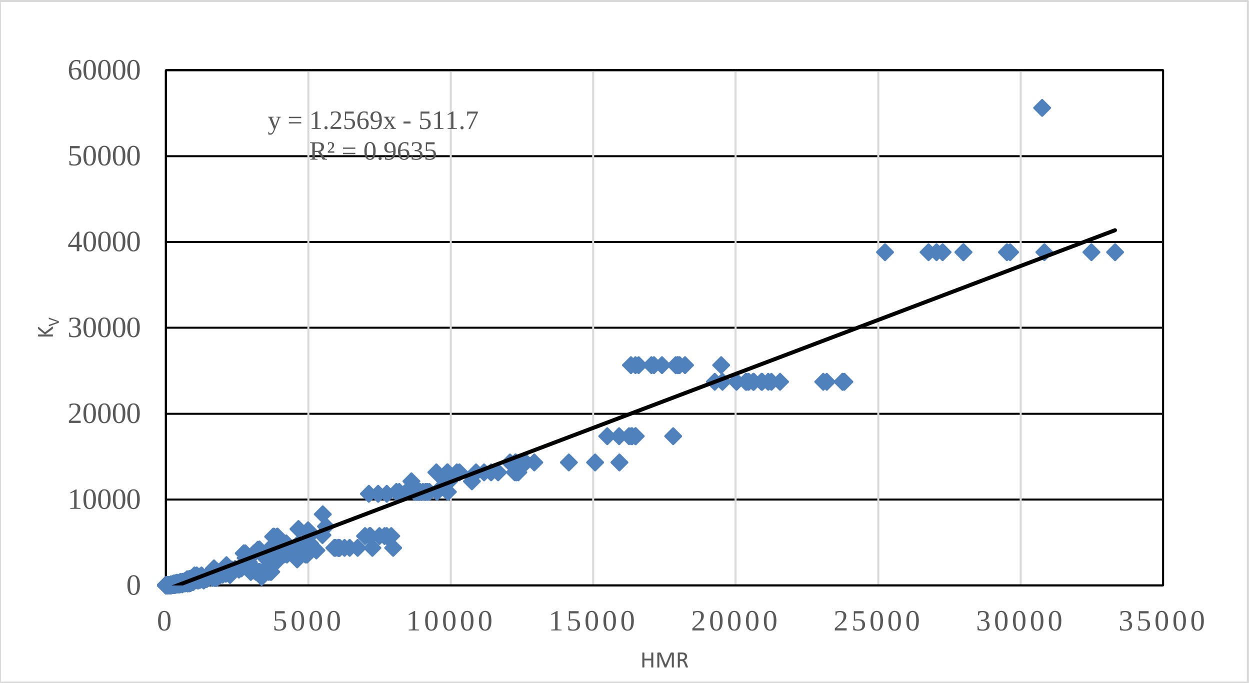


Figure 6. 8: The plot of Kv vs Kh and the developed new correlation coefficient is 0.906.

$$Kv = 1.256 * \frac{Kh}{PHIF} - 511.7 \tag{6.3}$$

$$\frac{K_v}{K_h} = \frac{1.256}{PHIF} - \frac{511.7}{Kh} \tag{6.4}$$

Table 6. 3:List of different correlations of Kv was studied by different authors including these studies

S.No	Correlation between K_v and K_h	R^2	Formation	Authors
1.	$\log Kv = 0.877 * \log Kh + 0.1963$	0.93	Arun limestone	(Atmadibrata et al., 1993)
2.	$Kv = 0.2997 * Kh^{1.0707}$	0.95	Niger- Delta sandstone	(Iheanacho et al., 2012)
3.	$Kv = 2.4484 * Kh$	0.69	Arbuckler formation	(Fazelalavi et al., 2015)
4.	$Kv = 0.5308 * Kh^{0.9707}$	0.83	Abu Rosh “C”sandstone	(Barakat & Nooh, 2017)
5.	$Kv = 0.7903 * Kh - 3.431$	0.86	Sandstone	(Shedid, 2019)
6.	$Kv = 1.256 * \frac{Kh}{PHIF} - 511.7$	96.3	Hugin sandstone formation	This study

Above equation 6.4 indicate that the anisotropy ratio Hugin sandstone formation permeability decreases with porosity. Hence the equation is important to determine before proper well spacing and pattern for suitable and efficient waterflood. Another important aspect accounted by Joshi (1991) for productivity index and fluid flow equation is influenced by vertical and horizontal permeability ratio. Joshi’s method for estimating productivity index in considerations of reservoir anisotropy.

Further investigation of permeability anisotropy needs to consideration of the relative distribution of permeability along the direction. We used the predicted K_v and K_h data of well #1 to study the nature of permeability distribution. Based on regression plot K_v vs K_h the values of vertical permeability are greater than horizontal permeability, this shows horizontal anisotropic. For this anisotropic case, the flow dominated was horizontal (radial) flow. The computed $\frac{K_v}{K_h}$ from equation 6.4 and plotted against depth to visualize whether permeability anisotropic varies systematically with depth in figure 6.10. Visualization of permeability anisotropic of well #1, The cross plot of K_v and K_h can represent the reservoir anisotropic if $\frac{K_v}{K_h}$ is greater than zero than the reservoir considered as vertical anisotropic whereas at any depth range if $\frac{K_v}{K_h}$ is bellow the X-axis the or less than zero, the reservoir is considered as horizontal anisotropic. K_v

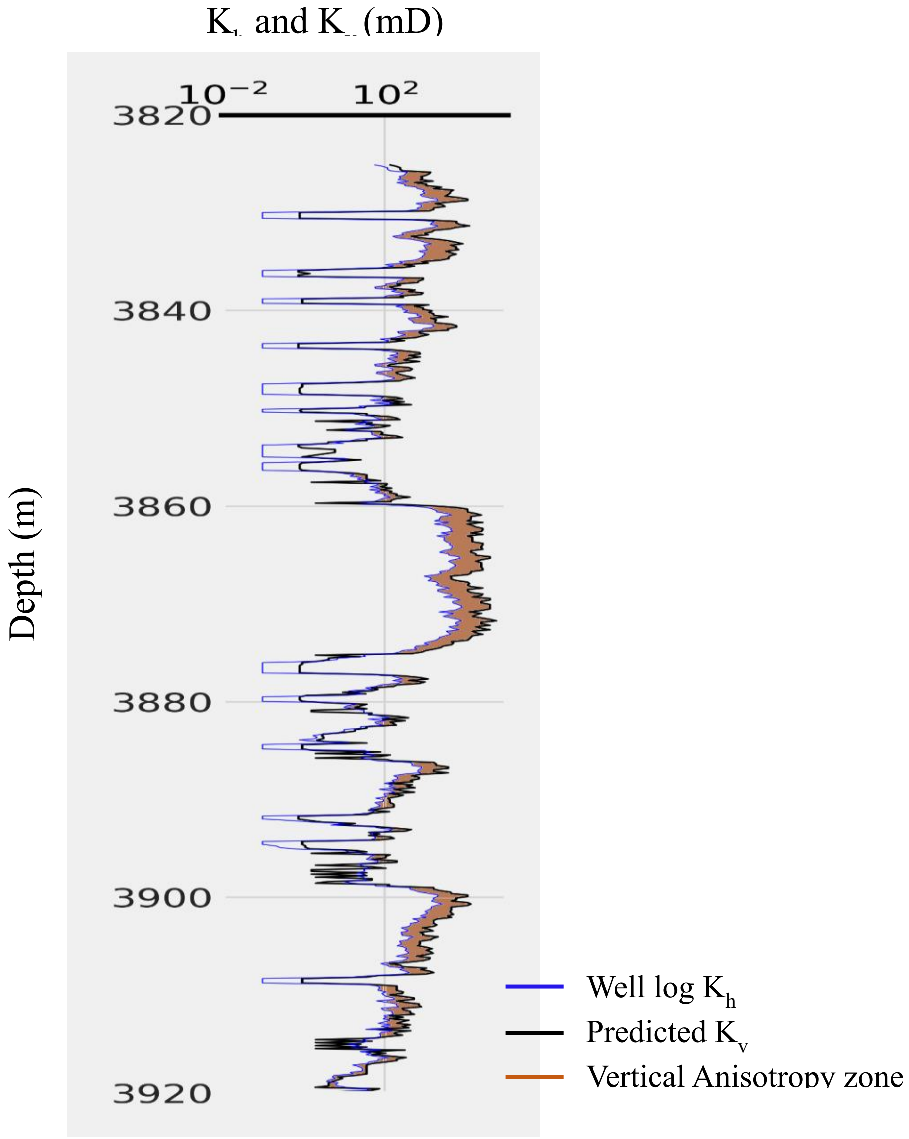


Figure 6. 10: Crossplot of vertical and horizontal permeability showing vertical anisotropy zone of well #1.

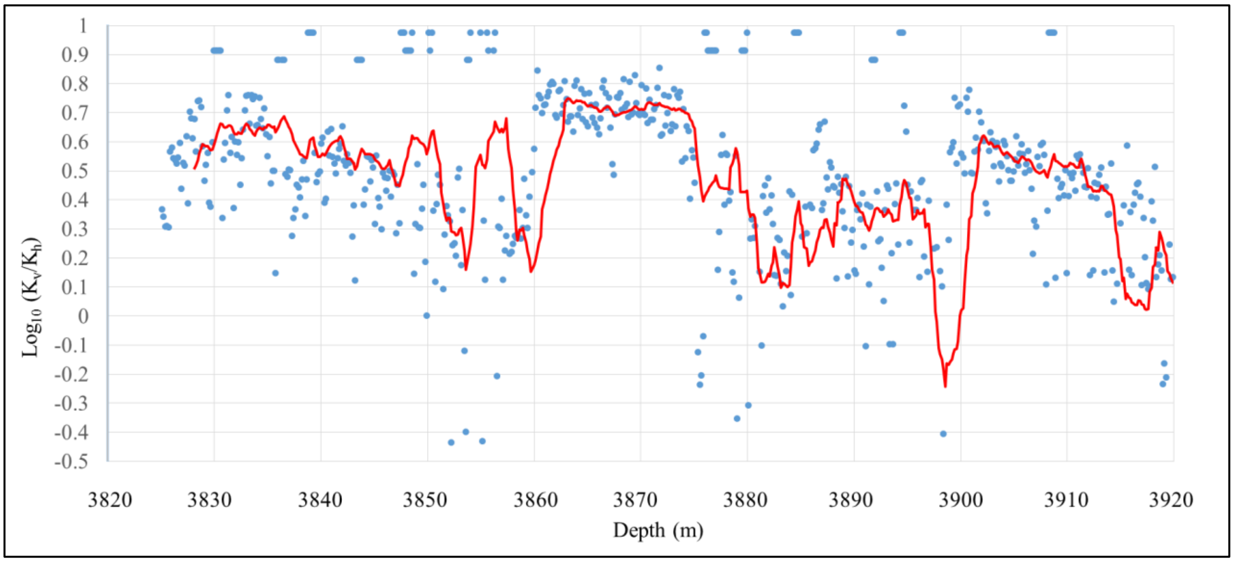


Figure 6. 10: A plot of K_v/K_h ratio vs depth for of Hugin formation sandstone over a depth range of 100 meters, to know whether horizontal permeability anisotropy varies systematically with depth.

CHAPTER- 7: CONCLUSIONS AND FUTURE WORK

For drive empirical correlation and to study reservoir anisotropy of Hugin sandstone formation. We applied data processing methodologies like data cleaning, logs smoothening, feature engineering, treating the outliers, etc. for implementing two shallows and three decision tree including two ensemble machine learning models (regression type problem supervised-learning) to the prediction of vertical permeability of sandstone reservoir. The five machine learning models were used to train two different training sets of two different cases. In the first case, no core data is available, to train the model in consideration of two well logs of data samples are 1560, where the target K_v is computed by Statoil and given for well #2 and well #3. In the second case, the core K_v is set as a target to train the model with 98 data samples. The ensemble random forest model exhibits good prediction over the shallow learning model. However shallow multi linear regression model performance was a fair and good model among all models for the prediction of K_v. Thus the RF is the best performing model the coefficient of determination is quite high 0.99. Apart from that, the stability of the model is also investigated using a cross-validation sampling technique before final prediction. The result indicates the shallow ANN model is less stable on the test data of well #3. However, the RF model was successfully tested on blind well #1 and validate with available field core data of the same well. Simultaneously the vertical permeability empirical model is developed and tested on core data. Thus the empirical model follows the trends of permeability. However, the empirical model is fair to estimate the extreme minimum and maximum values of permeability. Further, we develop the empirical correlation for vertical permeability of Hugin sandstone formation using the OLS (ordinary least square) linear regression model. In the study of permeability anisotropic nature in consideration of the vertical and horizontal variation of permeability along with direction on well #1, we found the reservoir is horizontal anisotropic. Apart from that new empirical correlation has been developed between permeability anisotropic ratio (K_v/K_h) and porosity, it can be used to estimate the anisotropic ratio from porosity core data. The developed correlation expresses useful application for Joshi's method for calculation for productivity index, to determine the proper well location (helps for suitable well flooding pattern). Overall this study has an impact on reservoir management and reservoir simulation for a better description of the reservoir.

Future work-

The zonation of the reservoir is an important aspect to identify and produce from multiple pay-zone. The zoning of well logs in consideration of all three well logs and core data.

1. Zonation using each different layer of anisotropic variation over depth.
2. The computational technique unsupervised method, in which the data points are analyzed for groups or clusters build within them, each cluster assigns each to a lithology.
3. Reservoir characterization

REFERENCES

Akkurt, R., Miller, M., Hodenfield, B., Pirie, I., Farnan, D., & Koley, M. (2019). Machine learning for well log normalization. *Proceedings - SPE Annual Technical Conference and Exhibition*, *2019-Septe*. https://doi.org/10.2118/196178-ms

Aliouane, L., Ouadfeul, S., Djarfour, N., & Boudella, A. (n.d.). *Petrophysical Parameters Estimation from Well-Logs Data Using Multilayer Perceptron*. 730–736.

Anderson, I. (1994). *Vertical Permeability from Resistivity Logs*.

Asquith, G. B. (Ed.). (1985). Use of Bulk Volume Water. In *Handbook of Log Evaluation Techniques for Carbonate Reservoirs* (Vol. 5, p. 0). American Association of Petroleum Geologists. https://doi.org/10.1306/Mth5446C3

Atmadibrata, R. R. M., Joenoes, S., & Oil, M. (1993). *Vertical-to-Horizontal Permeability Relationship: Arun Reservoir*. 365–371.

Ayan, C., Petricola, M., Knight, P., & Lalanne, B. (2007). An investigation of near-wellbore flow properties using sonic scanner measurements and interval pressure transient testing. *Proceedings - SPE Annual Technical Conference and Exhibition*, *4*, 2443–2452. https://doi.org/10.2118/110304-ms

Barakat, M. K., & Nooh, A. Z. (2017). Reservoir quality using the routine core analysis data of Abu Roash " C " in Badr. *Journal of African Earth Sciences, January 2019*. https://doi.org/10.1016/j.jafrearsci.2017.02.019

Basin, T. (2017). *Estimation of Shale Volume Using Gamma and Porosity Logs : Application to the Selected Gas Fields of Bangladesh*. *January*.

Chaki, S., Routray, A., & Mohanty, W. K. (2018). Well-Log and Seismic Data Integration for Reservoir Characterization: A Signal Processing and Machine-Learning Perspective. *IEEE Signal Processing Magazine*, *35*(2), 72–81. https://doi.org/10.1109/MSP.2017.2776602

Chen, H. C., & Chen, S. W. (n.d.). *Detection*.

Date, T. V. (2014). *Volve 15 / 9-F-1 B Petrophysical (static) well evaluation*. *February*.

Duchesne, M. J., & Gaillot, P. (2011). *Did you smooth your well logs the right way for seismic interpretation ?* *8*, 514–523. https://doi.org/10.1088/1742-2132/8/4/004

Fazelalavi, M., Formation, A., & Field, W. (2015). *Kv less than Kh - - - Group 3. June 2013*, 2–5.

Hou, J., Mekic, N., Quirein, J., Donderici, B., & Torres, D. (2016). Assessment of permeability anisotropy in anisotropic reservoirs with integrating multicomponent induction and conventional permeability logs. *Proceedings - SPE Annual Technical Conference and Exhibition*, *2016-Janua*. https://doi.org/10.2118/181449-ms

Iheanacho, P. C., Tiab, D., & Case, S. F. (2012). *SPE 163011 Vertical-Horizontal Permeability Relationships for Sandstone Reservoirs*. *1*, 1–8.

Jaiswal, J. K., & Das, R. (2018). Artificial neural network algorithms based nonlinear data analysis for forecasting in the finance sector. *International Journal of Engineering and Technology(UAE)*, *7*(4), 169–176. https://doi.org/10.14419/ijet.v7i4.10.20829

Johnson, W. W. (1994). Permeability determination from well logs and core data. *Proceedings of the Permian Basin Oil & Gas Recovery Conference*, 313–325. https://doi.org/10.2523/27647-ms

Li, H., & Misra, S. (2021). Robust machine-learning workflow for subsurface geomechanical characterization and comparison against popular empirical correlations. *Expert Systems with Applications*, *177*(March), 114942. https://doi.org/10.1016/j.eswa.2021.114942

Lim, J. S. (2005). Reservoir properties determination using fuzzy logic and neural networks from well data in offshore Korea. *Journal of Petroleum Science and Engineering*, *49*(3–4), 182–192. https://doi.org/10.1016/j.petrol.2005.05.005

Mabrouk, W. M. (2005). BVW as an indicator for hydrocarbon and reservoir homogeneity. *Journal of Petroleum Science and Engineering*, *49*(1–2), 57–62. https://doi.org/10.1016/j.petrol.2005.06.003

McCabe, K., & Horne, R. N. (2015). Estimating permeability anisotropy from downhole distributed temperature measurements. *Proceedings - SPE Annual Technical Conference and Exhibition*, *2015-Janua*, 3371–3390. https://doi.org/10.2118/174972-ms

Meyer, R. (2002). Anisotropy of sandstone permeability. *CREWES Reaserch Report*, *14*, 1–12. http://crewes.org/ForOurSponsors/ResearchReports/2002/2002-05.pdf

Mohaghegh, S., Balan, B., & Ameri, S. (1997). Permeability Determination From Well Log Data.

SPE Formation Evaluation, *12*(03), 170–174. https://doi.org/10.2118/30978-pa

Mohamad, A. M., & Hamada, G. M. (2017). Determination techniques of Archie's parameters: A, m and n in heterogeneous reservoirs. *Journal of Geophysics and Engineering*, *14*(6), 1358–1367. https://doi.org/10.1088/1742-2140/aa805c

Nordloh, V. A., Roubíčková, A., & Brown, N. (2020). Machine learning for gas and oil exploration. *Frontiers in Artificial Intelligence and Applications*, *325*, 3009–3016. https://doi.org/10.3233/FAIA200476

Rosenbrand, E., Fabricius, I. L., Fisher, Q., & Grattoni, C. (2015). Permeability in Rotliegend gas sandstones to gas and brine as predicted from NMR, mercury injection and image analysis. *Marine and Petroleum Geology*, *64*, 189–202. https://doi.org/10.1016/j.marpetgeo.2015.02.009

Salem, K. G. S. K. G., Abdulaziz, A. A. M. A. M., & Dahab, A. S. D. A. S. A. (2019). Prediction of hydraulic properties in carbonate reservoirs using artificial neural network. *Society of Petroleum Engineers - Abu Dhabi International Petroleum Exhibition and Conference 2018, ADIPEC 2018*. https://doi.org/10.2118/193007-ms

Sen, S., & Ganguli, S. S. (2019). Estimation of pore pressure and fracture gradient in volve field, Norwegian north sea. *Society of Petroleum Engineers - SPE Oil and Gas India Conference and Exhibition 2019, OGIC 2019*. https://doi.org/10.2118/194578-ms

Shedid, S. A. (2019). Prediction of vertical permeability and reservoir anisotropy using coring data. *Journal of Petroleum Exploration and Production Technology*, *0*(0), 0. https://doi.org/10.1007/s13202-019-0614-0

Sheng, J. J. (2008). Analytical steady-state solution of single-probe tests in a horizontal well and its application to estimate horizontal and vertical permeabilities. *SPE Reservoir Evaluation and Engineering*, *11*(3), 590–597. https://doi.org/10.2118/102659-pa

Singh, M., Makarychev, G., Mustapha, H., Voleti, D., Akkurt, R., Al Daghar, K., Mawlod, A. A., Al Marzouqi, K., Shehab, S., Maarouf, A., El Jundi, O., & Razouki, A. (2020). Machine learning assisted petrophysical logs quality control, editing and reconstruction. *Society of Petroleum Engineers - Abu Dhabi International Petroleum Exhibition and Conference 2020,* *ADIP 2020*. https://doi.org/10.2118/202977-ms

Sola, J., & Sevilla, J. (1997). Importance of input data normalization for the application of neural networks to complex industrial problems. *IEEE Transactions on Nuclear Science*, *44*(3 PART 3), 1464–1468. https://doi.org/10.1109/23.589532

Valent, M. B. (2019). *Estimation of permeability and e ff ective porosity logs using deep autoencoders in borehole image logs from the brazilian pre-salt carbonate*. 1–19.

Wang, B., Sharma, J., Chen, J., & Persaud, P. (2021). *Field Case Study*.

Wang, J., Zhang, L., Ge, K., Zhao, J., & Song, Y. (2020). Characterizing anisotropy changes in the permeability of hydrate sediment. *Energy*, *205*, 117997. https://doi.org/10.1016/j.energy.2020.117997

You, F. (2021). *Machine Learning application in Petrophysics Industry: A Sonic Log Synthesis prediction story | by Aboze Brain John Jnr | Towards Data Science*. 1–15. https://towardsdatascience.com/machine-learning-application-in-petrophysics-industry-a-sonic-log-synthesis-prediction-story-cf0ea54ffdad

Volve field report (2006) " \\statoil.net\unix_be\Project\volvepub\Well logs\05.PETROPHYSICALINTERPRETATION\2005_Sleipner _Øst_Hugin_Petrophysical_evaluation.doc "

Zagrebelnyy, E. V., Glushcenko, N. A., Martynov, M. E., Tsiklakov, A. M., Blinov, V. A., Weinheber, P., Karpekin, Y. A., Ezersky, D. M., & Bugakova, Y. S. (2017). Permeability anisotropy in the Thinly-bedded Pokurskaya formation from advanced wireline logs and formation testers. *Society of Petroleum Engineers - SPE Russian Petroleum Technology Conference 2017*. https://doi.org/10.2118/187760-ms

Zahaf, K., & Tiab, D. (2000). Vertical permeability from in situ horizontal measurements in shaly sand reservoirs. *Canadian International Petroleum Conference 2000, CIPC 2000*, *41*(8), 43–50. https://doi.org/10.2118/2000-006

Zhang, Y., & Ling, C. (2018). A strategy to apply machine learning to small datasets in materials science. *Npj Computational Materials*, *4*(1), 28–33. https://doi.org/10.1038/s41524-018-0081-z