Politecnico di Torino

Politecnico di Torino

Department of Environment, Land, and Infrastructure Engineering
Master of Science in Petroleum and Mining Engineering
A. y. 2023/2024

March 2024

Development of Predictive models of petrophysical properties using Machine Learning (ML) Algorithms on Well-log Data

Tutors:
Professor Laura Valentina Socco
Mr. Gabriel Sarantopoulos Bergamaschi

Candidate:
Amirhossein Akhondzadeh

Abstract

Reservoir characterization is an essential part of the hydrocarbon exploration and production workflow as it plays a significant role in our understanding of underground geological setting, fluid behaviour, and reservoir properties. Petrophysics is the discipline that allow to develop reservoir models and estimate their properties. Porosity and permeability assessment plays a significant role in reservoir characterization. Therefore, the main objective of this thesis is to predict and assess these parameters using machine learning algorithms.

The primary aim of this thesis is to estimate reservoir properties, including porosity, permeability, and saturation, utilizing well-log data. The focus lies on the diverse measurements obtained through various technologies. These measurements are indirect and necessitate interpretation processes grounded in assumed petrophysical models. Particularly in complex formations like carbonate reservoirs with intricate and heterogeneous porosity patterns, these models may lack full realism. Machine learning (ML) emerges as a promising approach in this context, offering multiple benefits. ML can enhance parameter estimation while also facilitating the extraction of static and dynamic properties, thereby reducing the need for a plethora of different measurements. For example, ML techniques can potentially obviate the necessity for NMR logs.

In the first part of the study, various predictive models, including Random Forest (RF), Gradient Boosting (GB), and K-Nearest Neighbor (K-NN), were developed. The NMR log provides a fairly accurate measurement of different types of porosity, such as total, effective, and free-fluid porosity. The basic well log data, such as gamma ray, resistivity, density, neutron porosity, acoustic slowness, and photoelectric factor, were used as predictors, while the Nuclear Magnetic Resonance (NMR) log was used as the target variable for the machine learning algorithms. Deterministic models, such as Timur-Coates, originally developed for sandstone reservoirs, were utilized to estimate permeability and saturation as functions of the three different porosity types estimated from the predictive models. Subsequently, the predicted parameters, including free fluid, effective, and total porosities, as well as permeability and saturation, were compared to the measured data for evaluation purposes.

In the second part of the study, the NMR log was bypassed, and permeability was directly used as the target variable for machine learning models. Other basic well-log data were utilized as predictors. This approach allows for the direct estimation of reservoir parameters using machine learning, which can be highly beneficial as it removes deterministic correlations, such as those found in the Timur-Coates model. In this part, a machine learning approach called Least Square Support Vector Regressor (LSSVR) was developed.

The outcomes of each part and the different models were evaluated using various regression metrics, such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the Coefficient of Determination (R2) values. The results indicate acceptable to highly accurate predictions for various cases, ranging from the test dataset to the entire dataset encompassing all five wells, as well as the evaluation of individual wells included in the training phase. However, when applying the models to a different well that was not included in the training phase, the results showed negative but still somewhat acceptable outcomes.

The Python code in the form of a Jupyter notebook is available on GitHub and is open to the public for enhanced accessibility and collaboration, facilitating potential future developments.

Acknowledgments

I would like to express my deep gratitude to my supervisor, Prof. Laura Valentina Socco, and my co-supervisor, Mr. Gabriel Sarantopoulos Bergamaschi, for their guidance and priceless mentorship during the duration of this thesis. Furthermore, I am immensely grateful to Vittoria De Pellegrini for generously sharing her codes, which proved invaluable in advancing the progress of my research.

I would like to extend my appreciation to the Brazilian National Agency for Petroleum, Natural Gas, and Biofuels (ANP, Agência Nacional de Petróleo) for providing access to the dataset and granting permission for its utilization in this research

I am profoundly thankful to my family and friend for their unwavering love and support. Their guidance has played a crucial role in shaping the individual I have become.

Finally, I would like to extend my heartfelt gratitude to the professors at the DIATI department for their unwavering dedication and tireless efforts in delivering exceptional education. Their commitment to excellence has truly enriched my learning experience and prepared me for future endeavors.

Table of Content

Abstract 2
Acknowledgments 4
Table of Content 5
List of Figures 8
List of Equations 14
List of Tables 15
Nomenclature 16
1 Introduction 19
2 Well-log Data Collection, Goal Definition, and Pre-processing 23
2.1 Data Collection and Pre-processing 23
2.2 Well-Log Data Information 26
2.3 Geological history 29
2.4 Geological formations 29
2.5 Geological features 30
2.6 Barra Velha Formation: 32
2.7 Locations and Well-log plots of different wells: 33
2.8 Visualizing Well Log Relationships and Distributions 42
2.8.1 Part One Correlations 45
2.8.2 Part Two Correlations 47
3 Methodologies and Concepts 48
3.1 Comprehensive Approaches to Permeability Estimation: Leveraging NMR and Well-Log Data Analysis in Predictive Modeling 48
3.2 Python library 50
3.3 Model Development and Optimization Techniques 51
3.3.1 Normalization 51
3.3.2 K-fold cross-validation (K-CV) 51
3.3.3 Hyperparameter tuning 53
3.3.4 Data Splitting 54
3.4 Part one (Prediction of Permeability from predicted NMR) 55
3.4.1 Defining Machine Learning (ML) Model: 57
3.4.1.1 Random forest: 57
3.4.1.2 Gradient Boosting: 57
3.4.1.3 k-nearest neighbor regression (k-NN): 58
3.4.2 Data Preparation and Splitting for NMR Fluid Prediction 59
3.4.3 Hyperparameter Optimization Across Models 60

3.4.4 Well-log data organisation (Test and all datasets) 62
3.4.5 The Timur-Coates model; Permeability & Saturation estimation 63
3.5 Part two (Prediction of Permeability directly from well-log data) 65
3.5.1 Defining Machine Learning (ML) Model: 66
3.5.2 Data Preparation and Splitting for Permeability Prediction 67
3.5.3 Hyperparameter Optimization Across Models 68
3.5.4 Well-log data organisation (Test and all datasets) 69
4 Methodology Application & Outcomes 70
4.1 Part one 70
4.1.1 Evaluation of different model performances using the test dataset 70
4.1.1.1 Scatter plots 70
4.1.1.2 Well-log Plots 72
4.1.2 Evaluation of different model performances using the entire dataset 75
4.1.2.1 Scatter plots 75
4.1.2.2 Well-log Plots 77
4.1.3 Evaluation of different model performances using the different dataset [1-BRSA-1116-RJS] (Not included in training-test) 80
4.1.3.1 Scatter plots 80
4.1.3.2 Well-log Plots 82
4.1.4 Evaluation of different model performances using single well [3-BRSA-1215-RJS] (Included in training-test) 85
4.1.4.1 Scatter plots 85
4.1.4.2 Well-log Plots 86
4.1.4.3 Well-log and Lab data plots 90
4.1.5 Evaluation of different model performances using single well [1-BRSA-1116-RJS] (Not included in training-test) 95
4.1.5.1 Well-log Plots 95
4.1.5.2 Well-log and Lab data plots 98
4.2 Part two 101
4.2.1 Evaluation of different model performances using the test dataset 101
4.2.1.1 Scatter Plot of Predicted and Measured permeability 101
4.2.1.2 Histogram of Predicted and Measured permeability 102
4.3 Using different regression metrics for model Evaluation 103
4.3.1 Regression metrics evaluation on the test data 104
4.3.2 Regression metrics evaluation on the entire dataset 106
4.3.3 Regression metrics evaluation on the different datasets (1-BRSA-1116-RJS) 109
5 Conclusion 111
Appendix 113
Appendix 01: (Histogram & Distributions plots) 113
Appendix 02: (Well-log plots) 120
Appendix 03: (Machine Learning Model approaches) 136

Appendix 04: (Developing Machine Learning Model) 142
References 143

List of Figures

Figure 1-1: Schematic representation of the constituents of a rock; Definition of different porosities....20

Figure 1-2: the general workflow of the thesis; First and Second Phases....22

Figure 2-1: General concept of first part of the predictive model; using basic well log data (input) vs NMR log data (target) to estimate reservoir characteristics from predicted NMR....24

Figure 2-2: General concept of second part of the predictive model; using basic well log data (input) vs permeability lab data (target) to estimate reservoir permeability directly from well log data....24

Figure 2-3: General workflow of the data collection and pre-processing; creation of a unique data frame of different wells in the Santos basin, for two parts of the thesis (Part 01 and Part 02)....25

Figure 2-4: Available well log data across all six wells; Gama Ray (GR), Resistivity at 90 inches (AT), Neutron Porosity (NPHI), Compressional Slowness (DTCO), Photoelectric Factor (PEFZ), Nuclear Magnetic Resonance (NMR)....28

Figure 2-5: General map of Santos Bason and Pre-salt reservoir location on the basin (Santos and Campos)....30

Figure 2-6: Stratigraphic chart of the rift and post-rift phases of the Santos Basin, modified (Link)....33

Figure 2-7: Map of six different wells [3-BRSA-944A-RJS, 1-BRSA-1116-RJS, 3-BRSA-1215-RJS, and 9-ITP-1-RJS, 3-EQNR-1-SPS and 3-EQNR-3-SPS] in fields of Itapù, Bùzios, Bacalhau (all located in the Santos Basin)....34

Figure 2-8: Map of three different fields (Itapù, Bùzios, Bacalhau); Six different wells are drilled into these three fields.....34

Figure 2-9: Pore Distribution in Well 9-ITP-RJS, Santos Basin: A close-up view showcasing the intricate pore structure and distribution.....35

Figure 2-10: Composite well-log of well 3-BRSA-944A-RJS (prediction depth interval: 5477 - 5712 m) vs Measured Depth. Track 1: Gamma-Ray (GR). Track 2: Induction Electric Resistivity logs. Investigation depths of 10 (AT10), 30 (AT30), and 90 (AT90) inches. Track 3: Formation Density (RHOZ), Neutron Porosity (NPHI), and Photoelectric Factor (PEFZ) logs. Track 4: Compressional Wave Slowness (DTCO). Track 5: Nuclear Magnetic Resonance Porosity logs. Total Porosity (NMRTOT), Effective Porosity (NMREFF), and Free Fluid (NMRFF). The light green highlight interval which is the reservoir rock.....36

Figure 2-11: Composite well-log of well 1-BRSA-1116-RJS (prediction depth interval: 5358 - 5560 m) vs Measured Depth. Track 1: Gamma-Ray (GR). Track 2: Induction Electric Resistivity logs. Investigation depths of 10 (AT10), 30 (AT30), and 90 (AT90) inches. Track 3: Formation Density (RHOZ), Neutron Porosity (NPHI), and Photoelectric Factor (PEFZ) logs. Track 4: Compressional Wave Slowness (DTCO). Track 5: Nuclear Magnetic Resonance Porosity logs. Total Porosity (NMRTOT), Effective Porosity (NMREFF), and Free Fluid (NMRFF). The light green highlight interval which is the reservoir rock.....37

Figure 2-12: Composite well-log of well 3-BRSA-1215-RJS (prediction depth interval: 5417 - 5650 m) vs Measured Depth. Track 1: Gamma-Ray (GR). Track 2: Induction Electric Resistivity logs. Investigation depths of 10 (AT10), 30 (AT30), and 90 (AT90) inches. Track 3: Formation Density (RHOZ), Neutron Porosity (NPHI), and Photoelectric Factor (PEFZ) logs. Track 4: Compressional Wave Slowness (DTCO). Track 5: Nuclear Magnetic Resonance Porosity logs. Total Porosity (NMRTOT), Effective Porosity (NMREFF), and Free Fluid (NMRFF). The light green highlight interval which is the reservoir rock.....38

Figure 2-13: Composite well-log of well 9-ITP-1-RJS (prediction depth interval: 5230 - 5590 m) vs Measured Depth. Track 1: Gamma-Ray (GR). Track 2: Induction Electric Resistivity logs. Investigation depths of 10 (AT10), 30 (AT30), and 90 (AT90) inches. Track 3: Formation Density (RHOZ), Neutron Porosity (NPHI), and Photoelectric Factor (PEFZ) logs. Track 4: Compressional Wave Slowness (DTCO). Track 5: Nuclear Magnetic Resonance Porosity logs. Total Porosity (NMRTOT), Effective Porosity (NMREFF), and Free Fluid (NMRFF). The light green highlight interval which is the reservoir rock.....39

Figure 2-14: Composite well-log of well 3-EQNR-1-SPS (prediction depth interval: 5910 - 6040 m) vs Measured Depth. Track 1: Gamma-Ray (GR). Track 2: Induction Electric Resistivity logs. Investigation depths of 10 (AT10), 30 (AT30), and 90 (AT90) inches. Track 3: Formation Density (RHOZ), Neutron Porosity (NPHI), and Photoelectric Factor (PEFZ) logs. Track 4: Compressional Wave Slowness (DTCO). Track 5: Nuclear Magnetic Resonance Porosity logs. Total Porosity (NMRTOT), Effective Porosity (NMREFF), and Free Fluid (NMRFF). The light green highlight interval which is the reservoir rock.....40

Figure 2-15: Composite well-log of well 3-EQNR-3-SPS (prediction depth interval: 6050 - 6315 m) vs Measured Depth. Track 1: Gamma-Ray (GR). Track 2: Induction Electric Resistivity logs. Investigation depths of 10 (AT10), 30 (AT30), and 90 (AT90) inches. Track 3: Formation Density (RHOZ), Neutron Porosity (NPHI), and Photoelectric Factor (PEFZ) logs. Track 4: Compressional Wave Slowness (DTCO). Track 5: Nuclear Magnetic Resonance Porosity logs. Total Porosity (NMRTOT), Effective Porosity (NMREFF), and Free Fluid (NMRFF). The light green highlight interval which is the reservoir rock.....41

Figure 2-16: Different Plots (KDE Plot, Histogram, and Strip) for RHOZ on the left and NPHI on the right side. Each of these plots for different properties (RHOZ & NPHI) show the distribution and comparison the range of the parameters for six different wells.....43

Figure 2-17: Different Plots (KDE Plot, Histogram, and Strip) for NMRFF on the left and NMREFF on the right side. Each of these plots for different properties (NMRFF & NMREFF) show the distribution and comparison the range of the parameters for six different wells.....44

Figure 2-18: Seaborn Pair plots of well: 3-BRSA-944A-RJS for different well log data (GR, AT90, RHOZ, NPHI, DTCO, PEFZ, and NMRTOT).....45

Figure 2-19: Seaborn Pair plots of well: 9-ITP-1-RJS for different well log data (GR, AT90, RHOZ, NPHI, DTCO, PEFZ, and NMRTOT)....46

Figure 2-20: Heatmap of correlation of well-log data and laboratory permeability related to the second part of the thesis....47

Figure 3-1: K-Fold Cross Validation Concept with K iterations across the Test & Train datasets....52

Figure 3-2: Definition of Predictors and Target for the process of Machine Learning (Part 01)....55

Figure 3-3: General workflow of the first part of thesis; train-test the data set contain five wells and prediction of permeability from output....56

Figure 3-4: Schematic of Random Forest (RF) decision tree Architecture.....57

Figure 3-5: Schematic of Gradient Boosting (GB) decision tree Architecture. 58

Figure 3-6: The illustration of K-Nearest Neighbor Regressor (K-NN). 59

Figure 3-7: Train-test splitting configuration 61

Figure 3-8: General description of wells, geology types, predictors, and target data (part 01)....62

Figure 3-9: Schematic representation of the constituents of a rock; Definition of different porosities; definition of BVM, BVI, CBW. 63

Figure 3-10:Definition of Predictors and Target for the process of Machine Learning (Part 02)....65

Figure 3-11: General workflow of the second part of thesis; train-test the data set contain five wells and prediction of permeability from output 66

Figure 3-12: The illustration of Support Vector Machine (SVM). 67

Figure 3-13: General description of wells, geology types, predictors, and target data (part 02)....69

Figure 4-1: Scatter plots of predicted versus measured NMR porosity, for the RF Model, applied to the test dataset of the original dataset contain whole five wells; Left plot: NMR Effective Porosity (m3/m3). Middle plot: NMR Free Fluid (m3/m3). Right plot: NMR Total Porosity (m3/m3). 70

Figure 4-2: Scatter plots of predicted versus measured NMR porosity, for the GB Model, applied to the test dataset of the original dataset contain whole five wells: Left plot: NMR Effective Porosity (m3/m3). Middle plot: NMR Free Fluid (m3/m3). Right plot: NMR Total Porosity (m3/m3). 70

Figure 4-3: Scatter plots of predicted versus measured NMR porosity, for the K-NN Model, applied to the test dataset of the original dataset contain whole five wells. Left plot: NMR Effective Porosity (m3/m3). Middle plot: NMR Free Fluid (m3/m3). Right plot: NMR Total Porosity (m3/m3). 71

Figure 4-4: Comparing the match between the predicted and measured NMR porosity, for the RF Model, across the Test Dataset Index of the original dataset contain whole five wells. Left track: Predicted and Measured NMR Effective Porosity (m3/m3). Middle track: Predicted and Measured NMR Free Fluid (m3/m3). Right track: Predicted and Measured NMR Total Porosity (m3/m3). 72

Figure 4-5: Comparing the match between the predicted and measured NMR porosity, for the GB Model, across the Test Dataset Index of the original dataset contain whole five wells. Left track: Predicted and Measured NMR Effective Porosity (m3/m3). Middle track: Predicted and Measured NMR Free Fluid (m3/m3). Right track: Predicted and Measured NMR Total Porosity (m3/m3). 73

Figure 4-6: Comparing the match between the predicted and measured NMR porosity, for the K-NN Model, across the Test Dataset Index of the original dataset contain whole five wells. Left track: Predicted and Measured NMR Effective Porosity (m3/m3). Middle track: Predicted and Measured NMR Free Fluid (m3/m3). Right track: Predicted and Measured NMR Total Porosity (m3/m3). 74

Figure 4-7: Scatter plots of predicted versus measured NMR porosity, for the RF Model, applied to the entire dataset of the original dataset contain whole five wells: Left plot: NMR Effective Porosity (m3/m3). Middle plot: NMR Free Fluid (m3/m3). Right plot: NMR Total Porosity (m3/m3). 75

Figure 4-8: Scatter plots of predicted versus measured NMR porosity, for the GB Model, applied to the entire dataset of the original dataset contain whole five wells: Left plot: NMR Effective Porosity (m3/m3). Middle plot: NMR Free Fluid (m3/m3). Right plot: NMR Total Porosity (m3/m3). 76

Figure 4-9: Scatter plots of predicted versus measured NMR porosity, for the K-NN Model, applied to the entire dataset of the original dataset contain whole five wells: Left plot: NMR Effective Porosity (m3/m3). Middle plot: NMR Free Fluid (m3/m3). Right plot: NMR Total Porosity (m3/m3). 76

Figure 4-10: Comparing the match between the predicted and measured NMR porosity, for the RF Model, across the Entire dataset index of the original dataset contain whole five wells. Left track: Predicted and Measured NMR Effective Porosity (m3/m3). Middle track: Predicted and Measured NMR Free Fluid (m3/m3). Right track: Predicted and Measured NMR Total Porosity (m3/m3). 77

Figure 4-11: Comparing the match between the predicted and measured NMR porosity, for the GB Model, across the Entire dataset index of the original dataset contain whole five wells. Left track: Predicted and Measured NMR Effective Porosity (m3/m3). Middle track: Predicted and Measured NMR Free Fluid (m3/m3). Right track: Predicted and Measured NMR Total Porosity (m3/m3). 78

Figure 4-12: Comparing the match between the predicted and measured NMR porosity, for the K-NN Model, across the Entire dataset index of the original dataset contain whole five wells. Left track: Predicted and Measured NMR Effective Porosity (m3/m3). Middle track: Predicted and Measured NMR Free Fluid (m3/m3). Right track: Predicted and Measured NMR Total Porosity (m3/m3). 79

Figure 4-13: Scatter plots of predicted versus measured NMR porosity, for the RF Model, applied to the different well (1-BRSA-1116-RJS) not included in training phase: Left plot: NMR Effective Porosity (m3/m3). Middle plot: NMR Free Fluid (m3/m3). Right plot: NMR Total Porosity (m3/m3)....80

Figure 4-14: Scatter plots of predicted versus measured NMR porosity, for the GB Model, applied to the different well (1-BRSA-1116-RJS) not included in training phase: Left plot: NMR Effective Porosity (m3/m3). Middle plot: NMR Free Fluid (m3/m3). Right plot: NMR Total Porosity (m3/m3)....81

Figure 4-15: Scatter plots of predicted versus measured NMR porosity, for the K-NN Model, applied to the different well (1-BRSA-1116-RJS) not included in training phase: Left plot: NMR Effective Porosity (m3/m3). Middle plot: NMR Free Fluid (m3/m3). Right plot: NMR Total Porosity (m3/m3)....81

Figure 4-16: Scatter plots of predicted versus measured NMR porosity, for the RF Model, applied to the different well (1-BRSA-1116-RJS) not included in training phase: Left plot: NMR Effective Porosity (m3/m3). Middle plot: NMR Free Fluid (m3/m3). Right plot: NMR Total Porosity (m3/m3)....81

Figure 4-17: Comparing the match between the predicted and measured NMR porosity, for the GB Model, applied to the different well (1-BRSA-1116-RJS): Left track: Predicted and Measured NMR Effective Porosity (m3/m3). Middle track: Predicted and Measured NMR Free Fluid (m3/m3). Right track: Predicted and Measured NMR Total Porosity (m3/m3). 83

Figure 4-18: Comparing the match between the predicted and measured NMR porosity, for the K-NN Model, applied to the different well (1-BRSA-1116-RJS): Left track: Predicted and Measured NMR Effective Porosity (m3/m3). Middle track: Predicted and Measured NMR Free Fluid (m3/m3). Right track: Predicted and Measured NMR Total Porosity (m3/m3). 84

Figure 4-19: Scatter plots of predicted versus measured NMR porosity, for the RF Model, applied to one of the well (3-BRSA-1215-RJS) included in training phase: Left plot: NMR Effective Porosity (m3/m3). Middle plot: NMR Free Fluid (m3/m3). Right plot: NMR Total Porosity (m3/m3). 85

Figure 4-20: Scatter plots of predicted versus measured NMR porosity, for the GB Model, applied to one of the well (3-BRSA-1215-RJS) included in training phase: Left plot: NMR Effective Porosity (m3/m3). Middle plot: NMR Free Fluid (m3/m3). Right plot: NMR Total Porosity (m3/m3). 85

Figure 4-21: Scatter plots of predicted versus measured NMR porosity, for the K-NN Model, applied to one of the well (3-BRSA-1215-RJS) included in training phase: Left plot: NMR Effective Porosity (m3/m3). Middle plot: NMR Free Fluid (m3/m3). Right plot: NMR Total Porosity (m3/m3). 86

Figure 4-22: Comparing the match between the predicted and measured NMR porosity (Effective, Free fluid Total), and Predicted Permeability & Saturation Using Timur-Coates Correlation, for the RF Model, applied to the well (3-BRSA-1215-RJS) included in training phase: (From left to right) 1st track: Predicted and Measured NMR Effective Porosity (m3/m3). 2nd track: Predicted and Measured NMR Free Fluid (m3/m3). 3rd track: Predicted and Measured NMR Total Porosity (m3/m3). 4th track: Predicted Effective, Free Fluid, Total Porosities (m3/m3), 5th track: Timur-Coates Predicted Permeability (mD), 6th track: Timur-Coates Predicted Irreducible water Saturation (%) 87

Figure 4-23: Comparing the match between the predicted and measured NMR porosity (Effective, Free fluid Total), and Predicted Permeability & Saturation Using Timur-Coates Correlation, for the GB Model, applied to the well (3-BRSA-1215-RJS) included in training phase: (From left to right) 1st track: Predicted and Measured NMR Effective Porosity (m3/m3). 2nd track: Predicted and Measured NMR Free Fluid (m3/m3). 3rd track: Predicted and Measured NMR Total Porosity (m3/m3). 4th track: Predicted Effective, Free Fluid, Total Porosities (m3/m3), 5th track: Timur-Coates Predicted Permeability (mD), 6th track: Timur-Coates Predicted Irreducible water Saturation (%) 88

Figure 4-24: Comparing the match between the predicted and measured NMR porosity (Effective, Free fluid Total), and Predicted Permeability & Saturation Using Timur-Coates Correlation, for the K-NN Model, applied to the well (3-BRSA-1215-RJS) included in training phase: (From left to right) 1st track: Predicted and Measured NMR Effective Porosity (m3/m3). 2nd track: Predicted and Measured NMR Free Fluid (m3/m3). 3rd track: Predicted and Measured NMR Total Porosity (m3/m3). 4th track: Predicted Effective, Free Fluid, Total Porosities (m3/m3), 5th track: Timur-Coates Predicted Permeability (mD), 6th track: Timur-Coates Predicted Irreducible water Saturation (%) 89

Figure 4-25: The reservoir interval of the well 3-BRSA-1215-RJS 90

Figure 4-26: Investigation of depth shift on log data in the well 3-BRSA-1215-RJS 91

Figure 4-27: Comparing the match between the predicted & laboratory measurement of permeability and porosity, Comparing scatter plot permeability-porosity of predicted and laboratory data, Histogram of Porosity and Permeability distribution of predicted and laboratory data, for the RF Model, applied to the well (3-BRSA-1215-RJS) included in training set (for Permeability Prediction Timur-Coates Correlation has been used): (From left to right) 1st track: Predicted and laboratory Permeability (mD). 2nd track: Predicted and Measured Porosity (%). 3rd track: Scatter plot of permeability-porosity of predicted and laboratory data (mD vs. %), 4th track: Histogram of Porosity of predicted and laboratory data (%), 5th track Histogram of Permeability of predicted and laboratory data (mD) 92

Figure 4-28: Comparing the match between the predicted & laboratory measurement of permeability and porosity, Comparing scatter plot permeability-porosity of predicted and laboratory data, Histogram of Porosity and Permeability distribution of predicted and laboratory data, for the GB Model, applied to the well (3-BRSA-1215-RJS) included in training set (for Permeability Prediction Timur-Coates Correlation has been used): (From left to right) 1st track: Predicted and laboratory Permeability (mD). 2nd track: Predicted and Measured Porosity (%). 3rd track: Scatter plot of permeability-porosity of predicted and laboratory data (mD vs. %), 4th track: Histogram of Porosity of predicted and laboratory data (%), 5th track Histogram of Permeability of predicted and laboratory data (mD) 93

Figure 4-29: Comparing the match between the predicted & laboratory measurement of permeability and porosity, Comparing scatter plot permeability-porosity of predicted and laboratory data, Histogram of Porosity and Permeability distribution of predicted and laboratory data, for the K-NN Model, applied to the well (3-BRSA-1215-RJS) included in training set (for Permeability Prediction Timur-Coates Correlation has been used): (From left to right) 1st track: Predicted and laboratory Permeability (mD). 2nd track: Predicted and Measured Porosity (%). 3rd track: Scatter plot of permeability-porosity of predicted and laboratory data (mD vs. %), 4th track: Histogram of Porosity of predicted and laboratory data (%), 5th track Histogram of Permeability of predicted and laboratory data (mD) 94

Figure 4-30: Comparing the match between the predicted and measured NMR porosity (Effective, Free fluid Total), and Predicted Permeability & Saturation Using Timur-Coates Correlation, for the RF Model, applied to a different well (3-BRSA-1116-RJS) not included in training phase: (From left to right) 1st track: Predicted and Measured NMR Effective Porosity (m3/m3). 2nd track: Predicted and Measured NMR Free Fluid (m3/m3). 3rd track: Predicted and Measured NMR Total Porosity (m3/m3). 4th track: Predicted Effective, Free Fluid, Total Porosities (m3/m3), 5th track: Timur-Coates Predicted Permeability (mD), 6th track: Timur-Coates Predicted Irreducible water Saturation (%) 95

Figure 4-31: Comparing the match between the predicted and measured NMR porosity (Effective, Free fluid Total), and Predicted Permeability & Saturation Using Timur-Coates Correlation, for the GB Model, applied to a different well (3-BRSA-1116-RJS) not included in training phase: (From left to right) 1st track: Predicted and Measured NMR Effective Porosity (m3/m3). 2nd track: Predicted and Measured NMR Free Fluid (m3/m3). 3rd track: Predicted and Measured NMR Total Porosity (m3/m3). 4th track: Predicted Effective, Free Fluid, Total Porosities (m3/m3), 5th track: Timur-Coates Predicted Permeability (mD), 6th track: Timur-Coates Predicted Irreducible water Saturation (%) 96

Figure 4-32: Comparing the match between the predicted and measured NMR porosity (Effective, Free fluid Total), and Predicted Permeability & Saturation Using Timur-Coates Correlation, for the K-NN Model, applied to a different well (3-BRSA-1116-RJS) not included in training phase: (From left to right) 1st track: Predicted and Measured NMR Effective Porosity (m3/m3). 2nd track: Predicted and Measured NMR Free Fluid (m3/m3). 3rd track: Predicted and Measured NMR Total Porosity (m3/m3). 4th track: Predicted Effective, Free Fluid, Total Porosities (m3/m3), 5th track: Timur-Coates Predicted Permeability (mD), 6th track: Timur-Coates Predicted Irreducible water Saturation (%) 97

Figure 4-33: Comparing the match between the predicted & laboratory measurement of permeability and porosity, Comparing scatter plot permeability-porosity of predicted and laboratory data, Histogram of Porosity and Permeability distribution of predicted and laboratory data, for the RF Model, applied to a different well (3-BRSA-1116-RJS) not included in training set (for Permeability Prediction Timur-Coates Correlation has been used): (From left to right) 1st track: Predicted and laboratory Permeability (mD). 2nd track: Predicted and Measured Porosity (%). 3rd track: Scatter plot of permeability-porosity of

predicted and laboratory data (mD vs. %), 4th track: Histogram of Porosity of predicted and laboratory data (%), 5th track Histogram of Permeability of predicted and laboratory data (mD)....98

Figure 4-34: Comparing the match between the predicted & laboratory measurement of permeability and porosity, Comparing scatter plot permeability-porosity of predicted and laboratory data, Histogram of Porosity and Permeability distribution of predicted and laboratory data, for the GB Model, applied to a different well (3-BRSA-1116-RJS) not included in training set (for Permeability Prediction Timur-Coates Correlation has been used): (From left to right) 1st track: Predicted and laboratory Permeability (mD). 2nd track: Predicted and Measured Porosity (%). 3rd track: Scatter plot of permeability-porosity of predicted and laboratory data (mD vs. %), 4th track: Histogram of Porosity of predicted and laboratory data (%), 5th track Histogram of Permeability of predicted and laboratory data (mD)....99

Figure 4-35: Comparing the match between the predicted & laboratory measurement of permeability and porosity, Comparing scatter plot permeability-porosity of predicted and laboratory data, Histogram of Porosity and Permeability distribution of predicted and laboratory data, for the GB Model, applied to a different well (3-BRSA-1116-RJS) not included in training set (for Permeability Prediction Timur-Coates Correlation has been used): (From left to right) 1st track: Predicted and laboratory Permeability (mD). 2nd track: Predicted and Measured Porosity (%). 3rd track: Scatter plot of permeability-porosity of predicted and laboratory data (mD vs. %), 4th track: Histogram of Porosity of predicted and laboratory data (%), 5th track Histogram of Permeability of predicted and laboratory data (mD)....100

Figure 4-36: Scatter plots of predicted versus measured Permeability (mD), for the LSSVR Model, applied to four wells [(3-BRSA-944A-RJS), (1-BRSA-1116-RJS), (9-ITP-RJS), (3-BRSA-1215-RJS)] included in training phase.....101

Figure 4-37: Comparative Histogram Analysis: LSSVR Predicted Permeability vs. Lab Permeability. Assessing the Frequency Distribution to Validate Model Performance....102

Figure 4-38: Random Forest Regression Metrics Histogram on Test Dataset: Comparative Analysis of Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Squared Error (MSE), and Coefficient of Determination (R^2) for Different Porosity Types.....105

Figure 4-39: Gradient Boosting Regression Metrics Histogram on Test Dataset: Comparative Analysis of Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Squared Error (MSE), and Coefficient of Determination (R^2) for Different Porosity Types.....105

Figure 4-40: K-Nearest Neighbour Regression Metrics Histogram on Test Dataset: Comparative Analysis of Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Squared Error (MSE), and Coefficient of Determination (R^2) for Different Porosity Types.....106

Figure 4-41: Random Forest Regression Metrics Histogram on Entire Dataset: Comparative Analysis of Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Squared Error (MSE), and Coefficient of Determination (R^2) for Different Porosity Types.....107

Figure 4-42: Gradient Boosting Regression Metrics Histogram on Entire Dataset: Comparative Analysis of Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Squared Error (MSE), and Coefficient of Determination (R^2) for Different Porosity Types.....107

Figure 4-43: K-Nearest Neighbour Regression Metrics Histogram on Entire Dataset: Comparative Analysis of Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Squared Error (MSE), and Coefficient of Determination (R^2) for Different Porosity Types.....108

Figure 4-44: Random Forest Regression Metrics Histogram on Different Dataset (Not Included in training phase): Comparative Analysis of Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Squared Error (MSE), and Coefficient of Determination (R^2) for Different Porosity Types.....109

Figure 4-45: Gradient Boosting Regression Metrics Histogram on Different Dataset (Not Included in training phase): Comparative Analysis of Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Squared Error (MSE), and Coefficient of Determination (R^2) for Different Porosity Types.....110

Figure 4-46: K-Nearest Neighbour Regression Metrics Histogram on Different Dataset (Not Included in training phase): Comparative Analysis of Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Squared Error (MSE), and Coefficient of Determination (R^2) for Different Porosity Types.....110

Figure A-1: Seaborn Pair plots of well: 1-BRSA-1116-RJS for different well log data (GR, AT90, RHOZ, NPHI, DTCO, PEFZ, and NMRTOT) 113

Figure A-2: Seaborn Pair plots of well: 3-BRSA-1215-RJS for different well log data (GR, AT90, RHOZ, NPHI, DTCO, PEFZ, and NMRTOT) 114

Figure A-3: Seaborn Pair plots of well: 3-EQNR-1-SPS for different well log data (GR, AT90, RHOZ, NPHI, DTCO, PEFZ, and NMRTOT).....115

Figure A-4: Seaborn Pair plots of well: 3-EQNR-3-SPS for different well log data (GR, AT90, RHOZ, NPHI, DTCO, PEFZ, and NMRTOT).....116

Figure A-5: Different Plots (KDE Plot, Histogram, and Strip) for GR on the left and AT90 on the right side. Each of these plots for different properties (GR & AT90) show the distribution and comparison the range of the parameters for six different wells.....117

Figure A-6: Different Plots (KDE Plot, Histogram, and Strip) for DTCO on the left and PEFZ on the right side. Each of these plots for different properties (DTCO & PEFZ) show the distribution and comparison the range of the parameters for six different wells. 118

Figure 0-7: Different Plots (KDE Plot, Histogram, and Strip) for NMRTOT show the distribution and comparison the range of the parameters for six different wells.....119

Figure A-8: Scatter plots of predicted versus measured NMR porosity, for the RF Model, applied to one of the well (3-BRSA-944A-RJS) included in training phase: Left plot: NMR Effective Porosity (m3/m3). Middle plot: NMR Free Fluid (m3/m3). Right plot: NMR Total Porosity (m3/m3). 120

Figure A-9: Scatter plots of predicted versus measured NMR porosity, for the GB Model, applied to one of the well (3-BRSA-944A-RJS) included in training phase: Left plot: NMR Effective Porosity (m3/m3). Middle plot: NMR Free Fluid (m3/m3). Right plot: NMR Total Porosity (m3/m3). 120

Figure A-10: Scatter plots of predicted versus measured NMR porosity, for the K-NN Model, applied to one of the well (3-BRSA-944A-RJS) included in training phase: Left plot: NMR Effective Porosity (m3/m3). Middle plot: NMR Free Fluid (m3/m3). Right plot: NMR Total Porosity (m3/m3). 121

Figure A-11: Comparing the match between the predicted and measured NMR porosity (Effective, Free fluid Total), and Predicted Permeability & Saturation Using Timur-Coates Correlation, for the RF Model, applied to the well (3-BRSA-944A-RJS) included in training phase: (From left to right) 1st track: Predicted and Measured NMR Effective Porosity (m3/m3). 2nd track: Predicted and Measured NMR Free Fluid (m3/m3). 3rd track: Predicted and Measured NMR

Total Porosity (m3/m3). 4th track: Predicted Effective, Free Fluid, Total Porosities (m3/m3), 5th track: Timur-Coates Predicted Permeability (mD), 6th track: Timur-Coates Predicted Irreducible water Saturation (%) *122*

Figure A-12: Comparing the match between the predicted and measured NMR porosity (Effective, Free fluid Total), and Predicted Permeability & Saturation Using Timur-Coates Correlation, for the GB Model, applied to the well (3-BRSA-944A-RJS) included in training phase: (From left to right) 1st track: Predicted and Measured NMR Effective Porosity (m3/m3). 2nd track: Predicted and Measured NMR Free Fluid (m3/m3). 3rd track: Predicted and Measured NMR Total Porosity (m3/m3). 4th track: Predicted Effective, Free Fluid, Total Porosities (m3/m3), 5th track: Timur-Coates Predicted Permeability (mD), 6th track: Timur-Coates Predicted Irreducible water Saturation (%) *123*

Figure A-13: Comparing the match between the predicted and measured NMR porosity (Effective, Free fluid Total), and Predicted Permeability & Saturation Using Timur-Coates Correlation, for the K-NN Model, applied to the well (3-BRSA-944A-RJS) included in training phase: (From left to right) 1st track: Predicted and Measured NMR Effective Porosity (m3/m3). 2nd track: Predicted and Measured NMR Free Fluid (m3/m3). 3rd track: Predicted and Measured NMR Total Porosity (m3/m3). 4th track: Predicted Effective, Free Fluid, Total Porosities (m3/m3), 5th track: Timur-Coates Predicted Permeability (mD), 6th track: Timur-Coates Predicted Irreducible water Saturation (%) *124*

Figure A-14: Comparing the match between the predicted & laboratory measurement of permeability and porosity, Comparing scatter plot permeability-porosity of predicted and laboratory data, Histogram of Porosity and Permeability distribution of predicted and laboratory data, for the RF Model, applied to the well (3-BRSA-944A-RJS) included in training set (for Permeability Prediction Timur-Coates Correlation has been used): (From left to right) 1st track: Predicted and laboratory Permeability (mD). 2nd track: Predicted and Measured Porosity (%). 3rd track: Scatter plot of permeability-porosity of predicted and laboratory data (mD vs. %), 4th track: Histogram of Porosity of predicted and laboratory data (%), 5th track Histogram of Permeability of predicted and laboratory data (mD) *125*

Figure A-15: Comparing the match between the predicted & laboratory measurement of permeability and porosity, Comparing scatter plot permeability-porosity of predicted and laboratory data, Histogram of Porosity and Permeability distribution of predicted and laboratory data, for the GB Model, applied to the well (3-BRSA-944A-RJS) included in training set (for Permeability Prediction Timur-Coates Correlation has been used): (From left to right) 1st track: Predicted and laboratory Permeability (mD). 2nd track: Predicted and Measured Porosity (%). 3rd track: Scatter plot of permeability-porosity of predicted and laboratory data (mD vs. %), 4th track: Histogram of Porosity of predicted and laboratory data (%), 5th track Histogram of Permeability of predicted and laboratory data (mD) *126*

Figure A-16: Comparing the match between the predicted & laboratory measurement of permeability and porosity, Comparing scatter plot permeability-porosity of predicted and laboratory data, Histogram of Porosity and Permeability distribution of predicted and laboratory data, for the K-NN Model, applied to the well (3-BRSA-944A-RJS) included in training set (for Permeability Prediction Timur-Coates Correlation has been used): (From left to right) 1st track: Predicted and laboratory Permeability (mD). 2nd track: Predicted and Measured Porosity (%). 3rd track: Scatter plot of permeability-porosity of predicted and laboratory data (mD vs. %), 4th track: Histogram of Porosity of predicted and laboratory data (%), 5th track Histogram of Permeability of predicted and laboratory data (mD) *127*

Figure A-17: Scatter plots of predicted versus measured NMR porosity, for the RF Model, applied to one of the well (9-ITP-RJS) included in training phase: Left plot: NMR Effective Porosity (m3/m3). Middle plot: NMR Free Fluid (m3/m3). Right plot: NMR Total Porosity (m3/m3). *128*

Figure A-18: Scatter plots of predicted versus measured NMR porosity, for the GB Model, applied to one of the well (9-ITP-RJS) included in training phase: Left plot: NMR Effective Porosity (m3/m3). Middle plot: NMR Free Fluid (m3/m3). Right plot: NMR Total Porosity (m3/m3). *128*

Figure A-19: Scatter plots of predicted versus measured NMR porosity, for the K-NN Model, applied to one of the well (9-ITP-RJS) included in training phase: Left plot: NMR Effective Porosity (m3/m3). Middle plot: NMR Free Fluid (m3/m3). Right plot: NMR Total Porosity (m3/m3). *129*

Figure A-20: Comparing the match between the predicted and measured NMR porosity (Effective, Free fluid Total), and Predicted Permeability & Saturation Using Timur-Coates Correlation, for the RF Model, applied to the well (9-ITP-RJS) included in training phase: (From left to right) 1st track: Predicted and Measured NMR Effective Porosity (m3/m3). 2nd track: Predicted and Measured NMR Free Fluid (m3/m3). 3rd track: Predicted and Measured NMR Total Porosity (m3/m3). 4th track: Predicted Effective, Free Fluid, Total Porosities (m3/m3), 5th track: Timur-Coates Predicted Permeability (mD), 6th track: Timur-Coates Predicted Irreducible water Saturation (%) *130*

Figure A-21: Comparing the match between the predicted and measured NMR porosity (Effective, Free fluid Total), and Predicted Permeability & Saturation Using Timur-Coates Correlation, for the GB Model, applied to the well (9-ITP-RJS) included in training phase: (From left to right) 1st track: Predicted and Measured NMR Effective Porosity (m3/m3). 2nd track: Predicted and Measured NMR Free Fluid (m3/m3). 3rd track: Predicted and Measured NMR Total Porosity (m3/m3). 4th track: Predicted Effective, Free Fluid, Total Porosities (m3/m3), 5th track: Timur-Coates Predicted Permeability (mD), 6th track: Timur-Coates Predicted Irreducible water Saturation (%) *131*

Figure A-22: Comparing the match between the predicted and measured NMR porosity (Effective, Free fluid Total), and Predicted Permeability & Saturation Using Timur-Coates Correlation, for the K-NN Model, applied to the well (9-ITP-RJS) included in training phase: (From left to right) 1st track: Predicted and Measured NMR Effective Porosity (m3/m3). 2nd track: Predicted and Measured NMR Free Fluid (m3/m3). 3rd track: Predicted and Measured NMR Total Porosity (m3/m3). 4th track: Predicted Effective, Free Fluid, Total Porosities (m3/m3), 5th track: Timur-Coates Predicted Permeability (mD), 6th track: Timur-Coates Predicted Irreducible water Saturation (%) *132*

Figure A-23: Comparing the match between the predicted & laboratory measurement of permeability and porosity, Comparing scatter plot permeability-porosity of predicted and laboratory data, Histogram of Porosity and Permeability distribution of predicted and laboratory data, for the RF Model, applied to the well (3-BRSA-944A-RJS) included in training set (for Permeability Prediction Timur-Coates Correlation has been used): (From left to right) 1st track: Predicted and laboratory Permeability (mD). 2nd track: Predicted and Measured Porosity (%). 3rd track: Scatter plot of permeability-porosity of predicted and laboratory data (mD vs. %), 4th track: Histogram of Porosity of predicted and laboratory data (%), 5th track Histogram of Permeability of predicted and laboratory data (mD) *133*

Figure A-24: Comparing the match between the predicted & laboratory measurement of permeability and porosity, Comparing scatter plot permeability-porosity of predicted and laboratory data, Histogram of Porosity and Permeability distribution of predicted and laboratory data, for the GB Model, applied to the well (3-BRSA-944A-RJS) included in training set (for Permeability Prediction Timur-Coates Correlation has been used): (From left to right) 1st track: Predicted and laboratory Permeability (mD). 2nd track: Predicted and Measured Porosity (%). 3rd track: Scatter plot of permeability-porosity of predicted

and laboratory data (mD vs. %), 4th track: Histogram of Porosity of predicted and laboratory data (%), 5th track Histogram of Permeability of predicted and laboratory data (mD) *134*

Figure A-25: Comparing the match between the predicted & laboratory measurement of permeability and porosity, Comparing scatter plot permeability-porosity of predicted and laboratory data, Histogram of Porosity and Permeability distribution of predicted and laboratory data, for the K-NN Model, applied to the well (3-BRSA-944A-RJS) included in training set (for Permeability Prediction Timur-Coates Correlation has been used): (From left to right) 1st track: Predicted and laboratory Permeability (mD). 2nd track: Predicted and Measured Porosity (%). 3rd track: Scatter plot of permeability-porosity of predicted and laboratory data (mD vs. %), 4th track: Histogram of Porosity of predicted and laboratory data (%), 5th track Histogram of Permeability of predicted and laboratory data (mD) *135*

Figure A-26: The illustration of Support Vector Machine (SVM). *139*

Figure A-27: The illustration of K-Nearest Neighbor Regressor (K-NN). *141*

List of Equations

Equation 3-1 ..... *63*
Equation 3-2 ..... *63*
Equation 3-3 ..... *64*
Equation 3-4 ..... *64*
Equation 3-5 ..... *64*
Equation 3-6 ..... *64*
Equation 3-7 ..... *64*
Equation 3-8 ..... *64*
Equation 4-1 ..... *103*
Equation 4-2 ..... *103*
Equation 4-3 ..... *103*
Equation 4-4 ..... *104*
Equation A-1 ..... *136*
Equation A-2 ..... *136*
Equation A-3 ..... *136*
Equation A-4 ..... *137*
Equation A-5 ..... *137*
Equation A-6 ..... *138*
Equation A-7 ..... *138*
Equation A-8 ..... *138*
Equation A-9 ..... *138*
Equation A-10 ..... *138*
Equation A-11 ..... *138*
Equation A-12 ..... *139*
Equation A-13 ..... *140*
Equation A-14 ..... *140*

List of Tables

Table 0-1: Random Forest (RF), Gradient Boosting (GB), and K-Nearest Neighbour Regression (K-NN) Metrics on Test Dataset: Comparative Analysis of Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Squared Error (MSE), and Coefficient of Determination (R^2) for Different Porosity Types. 104

Table 0-2: Random Forest (RF), Gradient Boosting (GB), and K-Nearest Neighbour Regression (K-NN) Metrics on Entire Dataset: Comparative Analysis of Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Squared Error (MSE), and Coefficient of Determination (R^2) for Different Porosity Types. 106

Table 0-3: Random Forest (RF), Gradient Boosting (GB), and K-Nearest Neighbour Regression (K-NN) Metrics on Different Dataset (Not Included in training phase): Comparative Analysis of Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Squared Error (MSE), and Coefficient of Determination (R^2) for Different Porosity Types. 109

Nomenclature

DLIS: Digital Log Interchange Standard

GR: Gamma Ray Log

AT: Resistivity Log

AT10: Resistivity Log 10 inches from borehole

AT30: Resistivity Log 30 inches from borehole

AT90: Resistivity Log 90 inches from borehole

RHOZ: Density Log

NPHI: Neutron Porosity Log

PEFZ: Photoelectric factor log

DTCO: compressional slowness log

NMR: nuclear Magnetic Resonance

ML: Machine Learning

RF: Random Forest

GB: Gradient Boosting

K-NN: K-Nearest Neighbor

SVM: Support Vector Machine

LSSVR: Least Square Support Vector Regressor

Pd: Pandas

Np: Numpy

Plt: Matplotlib

Sns: Seaborn

K-CV: K-fold Cross Validation

COA: Cuckoo Optimization Algorithm

MSE: Mean Square Error

RMSE: Root Mean Square Error

MAE: Mean Absolute Error

R_2: Coefficient of Determination

ARI: Ariri Formation

BVE: Barra Velha Formation

BRSA: Brazil

RJS: Rio De Janeiro

SPS: Sao Paolo

SDR: Schulmberger-Doll-Research

BVM: Bulk Volume Movalbe

BVI: Bulk Volume Irreducible

CBW: Clay Bound Water Volume

BF: Bound Fluid

BFE: Bound Fluid Effective

ϕ_e: Effective Porosity

ϕ_t: Total Porosity

FF: Free Fluid Porosity

NMRFF: Free Fluid Porosity

NMREFF: Effective Porosity

NMRTOT: Total Porosity

Swi_e: Effective Irreducible Water Saturation

Swi_t: Total Irreducible Water Saturation

a: Coefficient in Timur-Coates Model

b: Coefficient in Timur-Coates Model

c: Coefficient in Timur-Coates Model

K: Permeability

DN: The dataset containing N samples

X: Input feature matrix

Y: Output SOH values

Xi: The ith sample point

Xij: The jth feature value of ith sample point

Yi: Measured SOH value

yi: Predicted SOH value

W: Weight matrix

β: Output weight matrix

k: Time step

g(·): Activation function

ψ(·) Mapping from input space to feature space

K(·) Kernel function

1 Introduction

Well log data analysis is a crucial aspect of the exploration and evaluation of hydrocarbon reservoirs, providing valuable insights into the physical properties and characteristics of subsurface formations. Traditionally, this analysis has relied on manual interpretation of well log curves and logs, which can be time-consuming, subjective, and prone to errors (Lee, 2002). Furthermore, manual interpretation does not facilitate the integration of multiple datasets from different wells or locations, potentially leading to inconsistent or incomplete results (Bueno, 2014).

Machine learning techniques have emerged as a promising approach to automate and enhance well log data analysis in the petroleum industry. These computational methods have been applied for various purposes, including classification, regression, clustering, anomaly detection, feature extraction, and dimensionality reduction. For instance, demonstrated the integration of well log interpretations for lithofacies classification and permeability modeling through advanced machine learning algorithms, overcoming multicollinearity and enhancing predictive capabilities. Furthermore, machine learning methods have shown potential for predicting water saturation distribution in reservoirs, offering valuable insights for reservoir characterization and management (Al-Mudhafar, 2007).

Furthermore, machine learning techniques have been utilized for interval inversion approaches to improve the interpretation of well logs, reducing the harmful effects of data noise and enhancing the accuracy of well-logging inverse problems (Mihály Dobróka, 2016). Additionally, the application of machine learning algorithms has enabled the prediction of the effectiveness of radial jet drilling technology in various geological conditions, providing valuable insights for optimizing drilling operations and production rates (Aleksandr Kochnev, 2021) (Zhao Wang, 2022). also emphasized the quantitative evaluation of unconsolidated sandstone heavy oil reservoirs based on machine learning, highlighting the potential of machine learning for reservoir evaluation and production optimization (Zhao Wang, 2022).

The application of machine learning techniques in the petroleum industry has gained significant attention for predicting well-log parameters without the need for additional costs or well intervention. This approach has been facilitated by the utilization of diverse well-log datasets to train machine learning models, thereby enhancing the accuracy of predictive models. Notably, machine learning techniques such as Random Forest, Gradient Boosting, K-nearest Neighbour, and Support Vector Machine have been widely applied in the petroleum industry for this purpose. Additionally, artificial neural networks have been utilized to analyze and predict trends and patterns in well-log data.

In the previous study, machine learning models such as random forest and gradient boosting were utilized for the prediction of various well-log data (including porosity, sonic, and NMR) based on basic well-log data in the Santos Basin. Pellegrini's work demonstrated the effectiveness of these models in predicting well log data (Pellegrini, 2023).

As a recommendation for future research, Pellegrini suggested that using the predicted NMR values, it may be possible to calculate additional petrophysical data such as permeability. Building upon this recommendation, the present thesis aims to further develop the methodology and techniques

introduced by Pellegrini. The focus of this thesis is to explore the potential for utilizing predicted NMR data to estimate petrophysical properties, specifically permeability, through the application of machine learning models. In the first part of this thesis, the basic well-log data are used as input for a machine learning approach. The goal is to build a model that can predict other specific logs, such as NMR (Nuclear Magnetic Resonance), from the well-log data. NMR is a non-destructive testing technique that measures the magnetic properties of fluids in porous media. By analyzing the NMR signals, different porosities (total, effective, and free fluid porosities) can be estimated. These porosities are important reservoir parameters that affect the fluid flow and trapping behavior in the subsurface.

The application of machine learning techniques for predicting NMR porosity from well-log data has been a subject of significant interest in the petroleum industry. This approach involves training and testing the model on a large dataset of well-log data from different wells and locations to establish the relationship between input well-log data (e.g., resistivity, sonic, density) and the output NMR porosity. The trained model can then be utilized to predict NMR porosity for new well-log data without requiring additional measurements or assumptions.

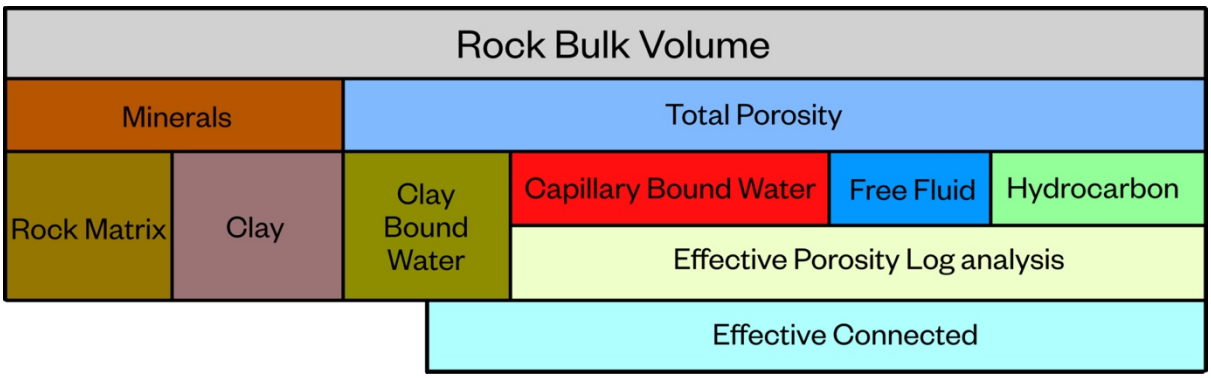

Figure 1-1: Schematic representation of the constituents of a rock; Definition of different porosities

Several studies have contributed to the advancement of machine learning applications in the petroleum industry, aligning with the objective of predicting reservoir parameters from well-log data. For instance, the synthesis of Nuclear Magnetic Resonance (NMR) outputs for clastic rocks using machine learning methods, emphasizing the significance of predicting NMR outputs for wells where NMR data is unavailable due to tool availability and logging costs (Rezaee, 2022). Additionally, it is demonstrated a machine learning accelerated approach to infer nuclear magnetic resonance porosity for a Middle Eastern carbonate reservoir, highlighting the reliability and consistency of the prediction models with low errors and high 'R' values (Ayyaz Mustafa, 2023). Furthermore, it is presented an adaptive boosting of the random forest algorithm for automatic petrophysical interpretation of well logs, showcasing the potential of machine learning techniques in petrophysical interpretation (Srivardhan, 2022).

The prediction of NMR porosity from well-log data will also lead to other reservoir parameters that can be derived from it, such as permeability and saturation. Permeability is a measure of how easily fluids can flow through a porous medium. Saturation is a measure of how much fluid is present in a pore space. These parameters are also essential for reservoir characterization and evaluation. By comparing

the predicted NMR porosity with lab data (RCAL), which is obtained by using reference cores or standard methods to measure porosity directly, acceptable values for permeability and saturation can be estimated. This will help to validate the accuracy and reliability of the machine learning model and its predictions.

The application of the least square support vector machine (LSSVM) for predicting permeability directly from basic well-log data has been a subject of significant interest in the petroleum industry. LSSVM, as a type of support vector machine (SVM), is capable of solving linear systems of equations by minimizing a quadratic objective function. It can effectively handle nonlinear and high-dimensional data by utilizing kernel functions to map the input data into a higher-dimensional feature space. LSSVM has been demonstrated to be effective and robust for various classification and regression problems in different domains.

The goal of applying LSSVM to predict permeability from well-log data is to develop a simple and fast method that does not require any additional measurements or assumptions about the reservoir properties. However, these methods have some limitations, such as requiring core samples for validation, being sensitive to noise and outliers, or being dependent on the quality and availability of well-log data.

The research consists of two main phases. In the first phase, well-log data of different wells are collected, pre-processed, and visualised. In the second phase, two different approaches to predict permeability from well-log data are applied and compared. The first approach uses nuclear magnetic resonance (NMR) log data. The NMR log data are used to estimate the pore size distribution and the effective porosity of the rock, and then to calculate the permeability using empirical models. The predicted permeability from NMR log data is compared with the measured permeability from laboratory tests on core samples taken from the same wells (figure 1-2).

The second approach uses basic well-log data, such as gamma ray, resistivity, density, and neutron porosity, and Photoelectric which are more widely available. The basic well-log data are used to estimate the shale volume, water saturation, and porosity of the rock, and then to calculate the permeability using least square support vector machine (LSSVM). The predicted permeability from basic well-log data is also compared with the measured permeability from laboratory tests on core samples (figure 1-2).

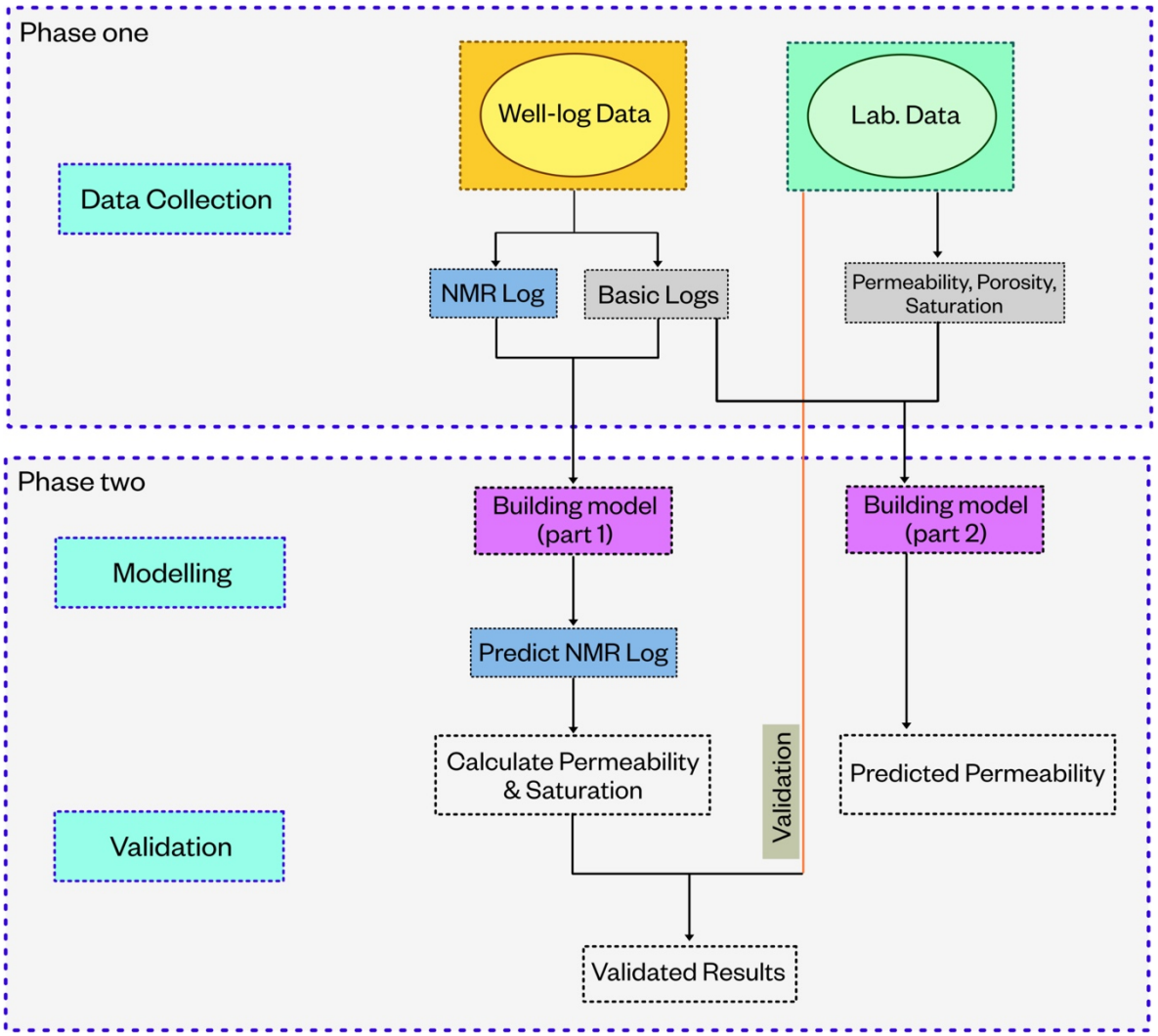

Figure 1-2: the general workflow of the thesis; First and Second Phases

2 Well-log Data Collection, Goal Definition, and Pre-processing

2.1 Data Collection and Pre-processing

The data collection process involved gathering well-log data from six different wells. Each well had various logging measurements, such as Gamma Ray (GR), Resistivity (AT), Density (RHOZ), Neutron Porosity (NPHI), Compressional Slowness (DTCO), Photo-electric Factor (PEFZ), and Nuclear Magnetic Resonance (NMR). The collected data was stored in a proprietary format called DLIS (Data Log Interchange Standard).

The initial stage in data preprocessing was to load the DLIS files from each well into a dedicated data collecting application. This tool allowed for the extraction of individual well-log measurements into distinct data frames. The following phase entailed removing any missing or incorrect data points. This was accomplished by recognizing and deleting any data points labeled as Nan (Not a Number) or falling outside of the expected range of values for the corresponding measurement.

After pre-processing, each well-log data set was represented by a distinct data frame, with each row representing a single depth measurement and each column representing a different logging measurement. The data frames were arranged depending on the well depths.

Reservoir Interval Extraction and Data Concatenation: The next step in data processing was to extract the depth range of the reservoir interval from each well-log data frame. The reservoir interval was defined as the stratigraphic unit that was the principal focus of the reservoir characterisation research. Once the reservoir interval was determined, the data points from each well-log data frame were retrieved and concatenated. This produced a single data frame containing all of the well-log measurements for the full reservoir interval.

Concatenating data frames required careful consideration of the depth scales utilized in the various well-log data sets. To ensure compatibility, the depth scales were thoroughly reviewed and changed as necessary to ensure that the data points were matched.

Visualizing Correlations and Data Distribution: A variety of visualizations were used to acquire a better understanding of the data and assure its viability for future analyses. Seaborn's pairplot tool was used to depict the correlation between all pairs of well-log data, revealing information about the correlations between various logging variables. In addition, histograms were created for each well-log measurement to examine data distribution and identify probable outliers or data abnormalities. These visualizations confirmed that the data was consistent, dependable, and appropriate for machine learning applications.

Final Data Frame and Predictive Modelling: After completing the data pre-processing and visualization stages, a comprehensive data frame was created, containing all essential well-log information for the reservoir interval. This data frame was used as the input for predictive machine learning (ML) models. The machine learning models were trained to infer connections between well-log values and reservoir parameters of interest, such as porosity, permeability, and hydrocarbon saturation. The trained ML models were subsequently utilized to predict these qualities at previously

unrecorded well locations, allowing for reservoir characterization outside the scope of accessible well data.

By properly collecting, pre-processing, displaying, and exploiting well-log data, a predictive ML model developed that precisely defines reservoir features during the whole interval. This model offers useful insights into reservoir management and optimization, allowing for more informed decisions about drilling, production, and reservoir development strategies (figures 2-1, 2-2, and 2-3).

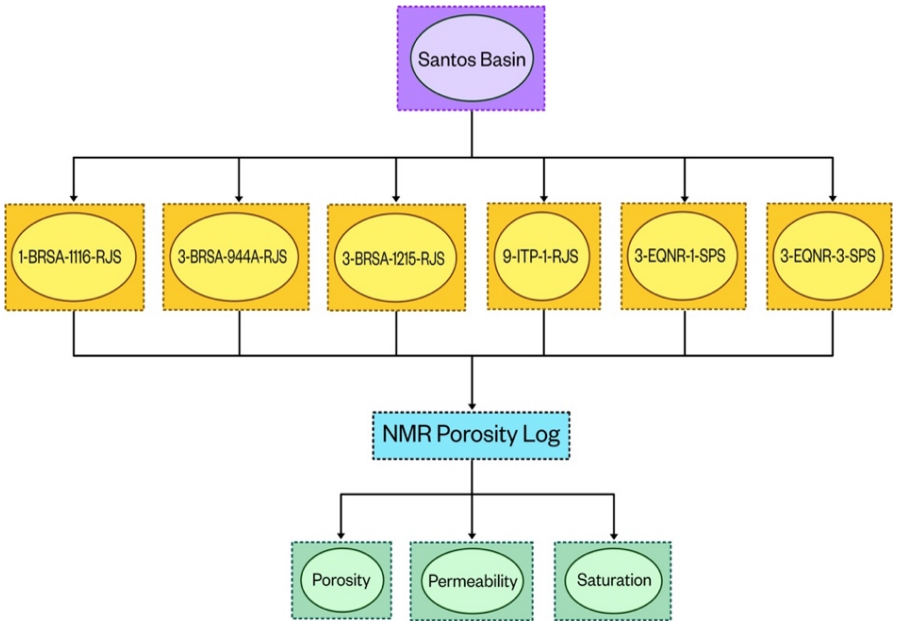

Figure 2-1: General concept of first part of the predictive model; using basic well log data (input) vs NMR log data (target) to estimate reservoir characteristics from predicted NMR

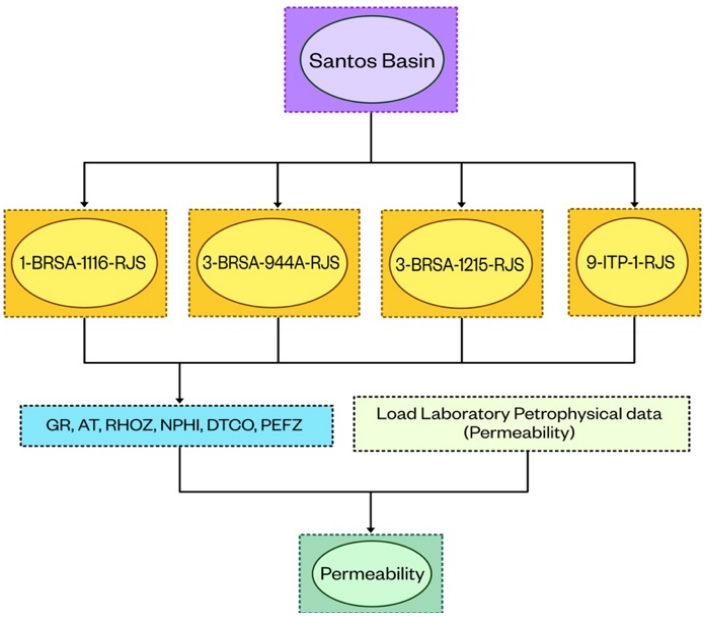

Figure 2-2: General concept of second part of the predictive model; using basic well log data (input) vs permeability lab data (target) to estimate reservoir permeability directly from well log data

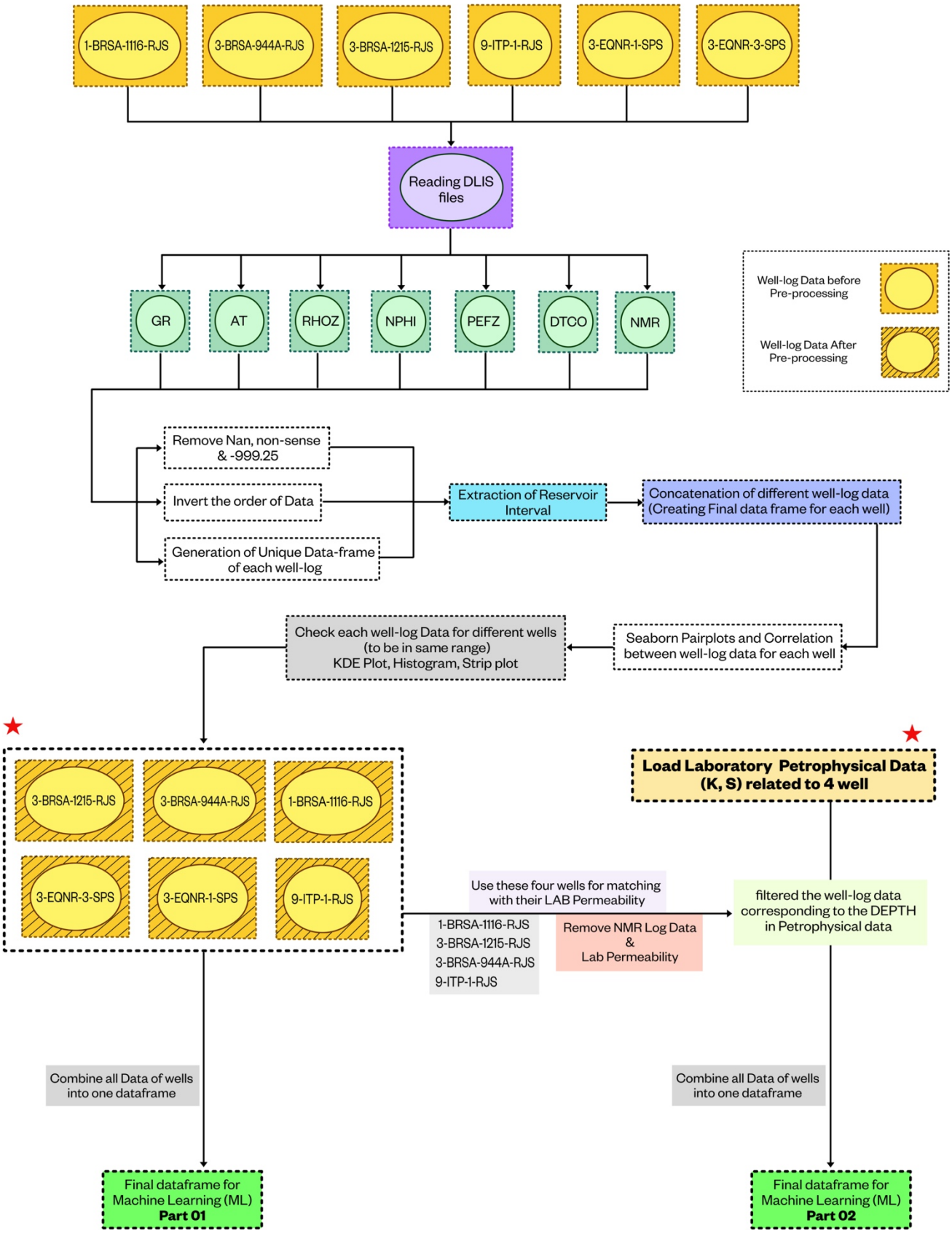

Figure 2-3: General workflow of the data collection and pre-processing; creation of a unique data frame of different wells in the Santos basin, for two parts of the thesis (Part 01 and Part 02)

2.2 Well-Log Data Information

Wireline logging is a crucial technique in the field of petroleum geology and exploration, offering valuable insights into subsurface formations and reservoir characteristics. Various wireline logging curves are used to measure different properties of the rock and fluid formations encountered during drilling and well completion. Each wireline logging curve provides specific information that is essential for understanding the geological and hydrocarbon potential of a well. Here, we'll briefly explain the importance of some key wireline logging curves, while more in-depth information can be found in open sources.

The Gamma-Ray log (GR), is a well logging instrument used in petroleum geology and reservoir evaluation that measures the natural radioactivity emitted by rock formations in a borehole, usually in API units. Its principal applications include detecting shaly zones with high radioactivity, which is useful for reservoir and formation appraisal. It also helps calculate shale volume, which is an important statistic for determining reservoir quality and hydrocarbon potential. Furthermore, the variable radioactivity levels associated with distinct clay minerals enable some degree of clay mineral identification in shales and clay-rich formations. The GR log is a valuable instrument in the oil and gas sector, providing critical data for drilling and production decisions in hydrocarbon exploration and production.

The Resistivity log (RES), is a crucial wireline logging curve used in petroleum geology and reservoir evaluation. Its major goal is to determine the resistivity of the geological formation, also known as apparent formation resistivity. This study considers inquiry depths of 10 inches (AT10), 30 inches (AT30), and 90 inches (AT90), which play an important role in analyzing the attributes of subsurface formations. The Resistivity log is important because it may be used to determine fluid saturation inside a formation. This is accomplished using the following principles:

- Resistivity and Formation Fluids: represent resistivity of a formation is highly dependent on the types and properties of fluids present within it. Different fluids, such as water and hydrocarbons, have distinct electrical conductivities. Water, being a good conductor, results in low resistivity values, while hydrocarbons, which are generally poor conductors, lead to high resistivity values.
- Depth Investigation: by measuring resistivity at specific depths, such as 10 inches, 30 inches, and 90 inches, geoscientists can obtain a vertical profile of resistivity variations within the wellbore. This allows for a more comprehensive assessment of the formation's properties at different invasion zones.
- Saturation Indication: the resistivity curves obtained at these depths provide valuable insights into the fluid saturation of the formation. Low resistivity values, as observed in the case of water-bearing zones, indicate the presence of water or brine, suggesting that these zones may not be ideal for hydrocarbon production. On the other hand, high resistivity values in the resistivity log maybe indicative of hydrocarbon-bearing zone, signifying the potential for the presence of oil or gas.

The Formation Density log, abbreviated as (RHOZ), is a crucial wireline logging tool that assesses subsurface formations by subjecting them to gamma ray bombardment. Its ability to gauge the degree of scattering and absorption of gamma rays provides a direct indication of the formation's bulk density,

revealing insights into lithology, lithological boundaries, and porosity. Particularly valuable in hydrocarbon exploration, this log helps identify hydrocarbon-bearing zones with distinctive bulk density characteristics, playing a pivotal role in well placement, reservoir assessment, and the overall success of oil and gas exploration and production operations.

The Neutron Porosity log, often denoted as (NPHI), is a critical tool in wireline logging that assesses subsurface formations by exposing them to a controlled source of fast neutrons. It gauges the formation's response to these neutrons, which in turn serves as a direct indicator of the formation's porosity. In essence, the log provides the porosity of the formation by evaluating the interaction of fast neutrons with the hydrogen atomic nuclei in the rock, providing essential data for reservoir characterization, hydrocarbon potential assessment, and wellbore stability decisions in the field of petroleum geology.

The Sonic log, specifically the Compressional Wave Slowness (DTCO), is a critical component of wireline logging in the petroleum industry. It serves as a valuable tool for gathering essential information about subsurface formations. In particular, it provides measurements related to the slowness (inverse of seismic velocity), and attenuation, of refracted acoustic waves that traverse the formation and reach the wellbore wall.

The Photo-Electric Factor log, often abbreviated as (PEFZ), is a significant wireline logging tool used in geological and petrophysical evaluations of subsurface formations. It operates by measuring the photo-electric absorption, which refers to the absorption of gamma rays under a specific threshold by the formation. This absorption response provides a highly valuable indicator of the formation's lithology, helping geologists and petrophysicists distinguish and characterize different rock types and mineral compositions within the wellbore. The PEFZ log's capacity to discern lithological variations is instrumental in making informed decisions related to reservoir mapping, well completion, and the overall understanding of geological formations in the context of the oil and gas industry.

Nuclear Magnetic Resonance (NMR), logging is a well logging method that measures the induced magnet moment of hydrogen nuclei (protons) contained within the fluid-filled pore space of porous media (reservoir rocks). The NMR signal is a function of the number of protons, the distribution of pore sizes, and the mobility of the fluids in the pore space. NMR logging is a powerful tool for reservoir characterization because it can provide information about the following:

Total porosity: The total porosity of a rock is the volume of all the pores in the rock divided by the total volume of the rock. NMR logging can be used to measure total porosity by measuring the total amount of hydrogen in the pore space and the measure is independent from the matrix properties.

Pore size distribution: The pore size distribution of a rock is the distribution of the sizes of the pores in the rock. NMR logging can be used to measure pore size distribution by measuring the relaxation times of the hydrogen protons in the pore space. Relaxation times are shorter for smaller pores and longer for larger pores.

Fluid type identification: NMR logging can be used to identify the types of fluids in the pore space by measuring the chemical shift of the hydrogen protons. The characteristic relaxation time of the NMR signal and is dependent on the type of molecule that the hydrogen proton is bonded to.

Fluid mobility: The fluid mobility in the pore space is a measure of how easily the fluids can flow through the pores. NMR logging can be used to measure fluid mobility by measuring the diffusion of the hydrogen protons in the pore space. Diffusion is a process by which hydrogen protons move from areas of high concentration to areas of low concentration.

The information provided by NMR logging can be used to improve the understanding of reservoir properties, fluid distributions, and optimize hydrocarbon recovery strategies. For example, the total porosity and pore size distribution can be used to estimate the reservoir's permeability and irreducible water saturation. The fluid type identification can be used to identify zones that are likely to produce oil or gas. And the fluid mobility can be used to identify zones that are likely to be productive and zones that are likely to be bypassed by fluids.

Here are some specific examples of how NMR logging is used in petroleum geology and reservoir characterization:

- Identifying and characterizing different porosity types : NMR logging can be used to identify and characterize different porosity types, such as intergranular pores, vugs, and fractures. This information can be used to understand the reservoir's flow characteristics and to identify potential drilling and production problems.
- Estimating permeability: Permeability is a measure of how easily fluids can flow through a rock. NMR logging can be used to estimate permeability by measuring the pore size distribution and the fluid mobility in the pore space.

Here is the available well log data across six different wells (figure 2-4):

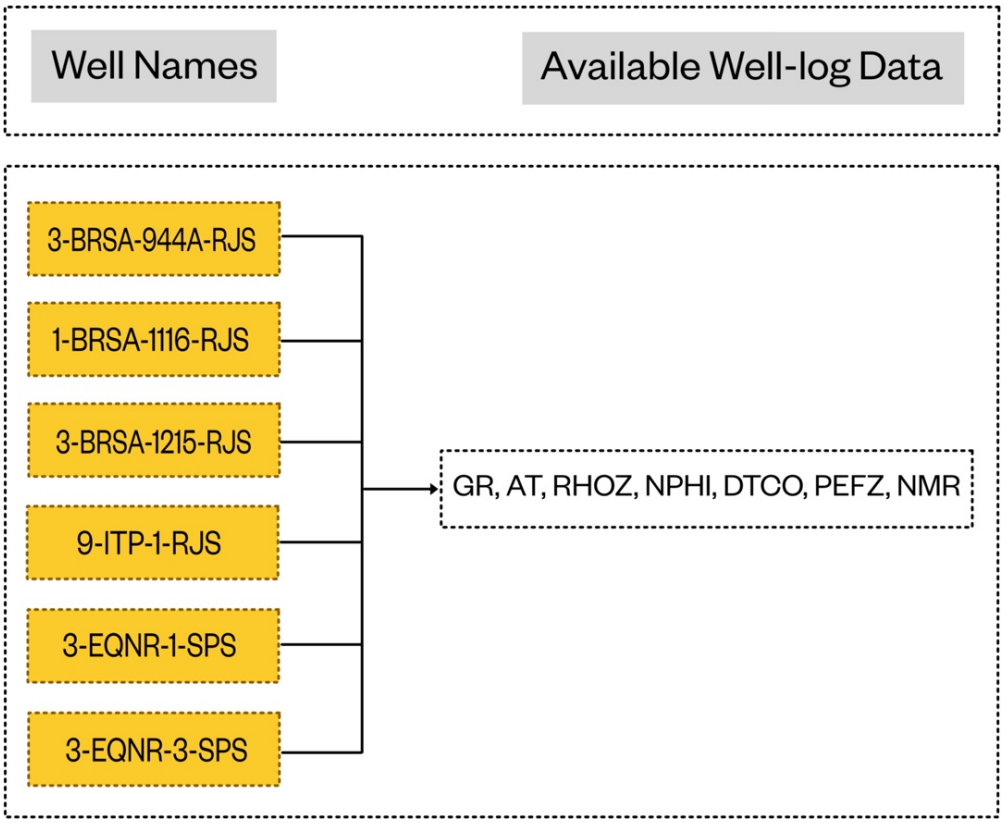

Figure 2-4: Available well log data across all six wells; Gama Ray (GR), Resistivity at 90 inches (AT), Neutron Porosity (NPHI), Compressional Slowness (DTCO), Photoelectric Factor (PEFZ), Nuclear Magnetic Resonance (NMR)

2.3 Geological history

The Santos Basin is bounded to the west by the Serra do Mar mountain range, and to the east by the São Paulo Plateau. The basin floor is divided into two main provinces: the Santos High, which is a shallow-water region, and the Santos Deep, which is a deeper-water region. Here's a summary of the key points you've highlighted:

Formation During Gondwana Breakup: The Santos Basin formed during the early Jurassic period as a result of the breakup of the supercontinent Gondwana. This geological event separated what are now the continents of South America and Africa.

Gondwana's Composition: Gondwana was a vast landmass that included present-day South America, Africa, India, Australia, and Antarctica. As it began to fragment, rift valleys formed, and the Santos Basin developed within one of these rift valleys.

Geographical Boundaries: The Santos Basin is geographically bounded by the Serra do Mar mountain range to the west and the São Paulo Plateau to the east, giving it distinct geographical boundaries.

Basin Floor Divisions: The basin floor of the Santos Basin can be divided into two primary provinces:

Santos High: This region is characterized as a shallow-water area.

Santos Deep: The Santos Deep is situated in deeper-water areas, likely representing greater depths and different geological features.

2.4 Geological formations

The Santos Basin is home to a variety of geological era's, including:

- Basement: The basement is composed of metamorphic and igneous rocks that formed during the Precambrian era. These rocks are located deep beneath the sedimentary rocks of the Santos Basin.
- Jurassic: The Jurassic section is composed of sandstones, shales, and limestones that were deposited in a variety of environments, including rivers, lakes, and deltas. These rocks were deposited during the early stages of the formation of the Santos Basin.
- Cretaceous: The Cretaceous section is composed of a thick layer of salt, followed by a sequence of sandstones and shales that were deposited in a marine environment. The salt layer was deposited during a period of restricted marine circulation, and the sandstones and shales were deposited after the salt layer was deposited.
- Paleogene: The Paleogene section is composed of a sequence of sandstones, shales, and limestones that were deposited in a variety of environments, including rivers, lakes, and deltas. These rocks were deposited after the salt layer was deposited and the marine circulation was restored.

- Neogene: The Neogene section is composed of a sequence of sandstones, shales, and limestones that were deposited in a marine environment. These rocks were deposited during the most recent geological period.

2.5 Geological features

The Santos Basin is also home to several important geological features, including:

Pre-salt reservoirs:

reservoir: Pre-salt reservoirs in the Santos Basin are located beneath a thick layer of salt that was deposited during the Cretaceous period. This salt layer acted as a seal, preserving organic-rich sediments deposited before and during its formation.

Hydrocarbon Potential: These reservoirs are known to contain some of the largest oil and gas reserves in the world. The hydrocarbons within these reservoirs are typically of high quality, with light crude oil being a common find.

Challenges: Developing pre-salt reservoirs is a formidable task due to their extreme depths, which can reach several kilometres beneath the seafloor. Drilling through the thick salt layer is technically challenging, as it requires specialized equipment and expertise to avoid complications like well instability and pressure imbalances.

Figure 2-5: General map of Santos Bason and Pre-salt reservoir location on the basin (Santos and Campos)

Salt diapirs:

Role in Trap Formation: Salt diapirs are of particular interest to the oil and gas industry because they can create structural traps for hydrocarbons. As the salt moves upward, it deforms the surrounding sedimentary rocks, causing them to fold and fracture. This deformation can create anticlines, domes, and other structural traps that are conducive to the accumulation of oil and gas.

Preservation of Hydrocarbons: The impermeable nature of salt layers can also contribute to the preservation of hydrocarbons. As the salt rises and seals off porous reservoir rocks, it prevents the escape of hydrocarbons, allowing them to accumulate over geological timeframes.

Common in Santos Basin: Salt diapirs are a common geological feature in the Santos Basin. Their presence is closely associated with the extensive salt layers found in this basin. These diapirs are often interconnected with other geological elements, such as faults and stratigraphic traps, further enhancing their importance in hydrocarbon exploration.

Geological Challenges: While salt diapirs provide opportunities for hydrocarbon exploration, they also present geological challenges. The complex deformation patterns associated with salt movement can result in reservoir heterogeneity, making accurate reservoir modeling and drilling planning crucial for success.

Grabens:

Grabens are structures that form when the Earth's crust is pulled apart. Grabens are common in the Santos Basin, and they have played an important role in the formation of the pre-salt reservoirs.

Formation Mechanism: Grabens are geological structures that form when the Earth's crust is subjected to extensional tectonic forces, causing it to be pulled apart. This stretching of the crust results in the creation of elongated, down-dropped blocks called grabens, often bounded by fault systems.

Role in Pre-salt Formation: Grabens have played a significant role in the formation of pre-salt reservoirs in the Santos Basin. These structural depressions provided ideal settings for the accumulation of sediments, including the organic-rich material that eventually became source rocks for hydrocarbons.

Accumulation of Sediments: Within the grabens, sedimentary layers accumulated over geological time. These layers often include the source rocks that generated the hydrocarbons, as well as reservoir rocks that can trap oil and gas.

Structural Traps: The faulted boundaries of grabens can create structural traps for hydrocarbons. As sediments filled the grabens, faulting could create both stratigraphic and structural traps, which are favorable conditions for the accumulation and retention of oil and gas.

Interaction with Salt Diapirs: Grabens in the Santos Basin are often interconnected with other geological elements, such as salt diapirs and fault systems. These interactions can further enhance their importance in hydrocarbon exploration. For example, salt diapirs can rise within grabens, creating additional structural complexity and trapping opportunities.

Geological Significance: Grabens are essential geological features that have influenced the basin's sedimentary architecture and hydrocarbon potential. Their presence and interaction with other

structural elements have contributed to the unique geological setting of the pre-salt play in the Santos Basin.

Challenges and Opportunities: While grabens provide opportunities for hydrocarbon exploration, they also present geological challenges. Understanding the complex deformation patterns associated with grabens is crucial for accurate reservoir modelling, drilling planning, and resource assessment in the basin.

2.6 Barra Velha Formation:

The Barra Velha Formation is a geological formation located in the southern region of Brazil, specifically in the Santos Basin offshore the state of São Paulo. This formation is of significant interest in the field of petroleum geology and exploration due to its hydrocarbon potential.

Here are some key points about the Barra Velha Formation:

Stratigraphy: The Barra Velha Formation is a part of the post-rift section within the Santos Basin. It is part of the late Cenozoic sedimentary sequence, primarily composed of clays, silts, sands, carbonate layers.

Hydrocarbon Potential: The Santos Basin, including the Barra Velha Formation, is known for its substantial hydrocarbon reserves. The formation is associated with both oil and gas deposits. Exploratory drilling and seismic studies have been conducted to assess the hydrocarbon potential of the region.

Reservoir Characteristics: The Barra Velha Formation features reservoir rocks with varying porosity and permeability. Understanding the porosity and fluid distribution within these reservoir rocks is critical for evaluating their production potential. The use of wireline logs like the Nuclear Magnetic Resonance (NMR) log, among others, is essential for characterizing these reservoirs.

Reservoir Management: The detailed knowledge of the Barra Velha Formation's lithology, porosity, and fluid content, obtained through wireline logging and core analysis, guides reservoir management decisions. This information aids in optimizing well placement, drilling strategies, and production schemes.

Exploration and Production: As oil and gas companies seek to expand their exploration efforts and increase production in the Santos Basin, the Barra Velha Formation remains a focal point for geological and petrophysical analysis. This includes using advanced logging tools like NMR to refine reservoir understanding and improve hydrocarbon recovery strategies.

the Barra Velha Formation in the Santos Basin, Brazil, is a geological formation of great significance in the context of hydrocarbon exploration and production. Advanced logging techniques, such as the Nuclear Magnetic Resonance (NMR) log, are vital for characterizing the reservoir properties within this formation and optimizing the extraction of oil and gas resources. It plays a central role in enhancing the petroleum industry's capacity to make informed decisions regarding the Santos Basin's potential as a valuable hydrocarbon reservoir. The stratigraphical information of Santos Basin has been shown in figure 2-6.

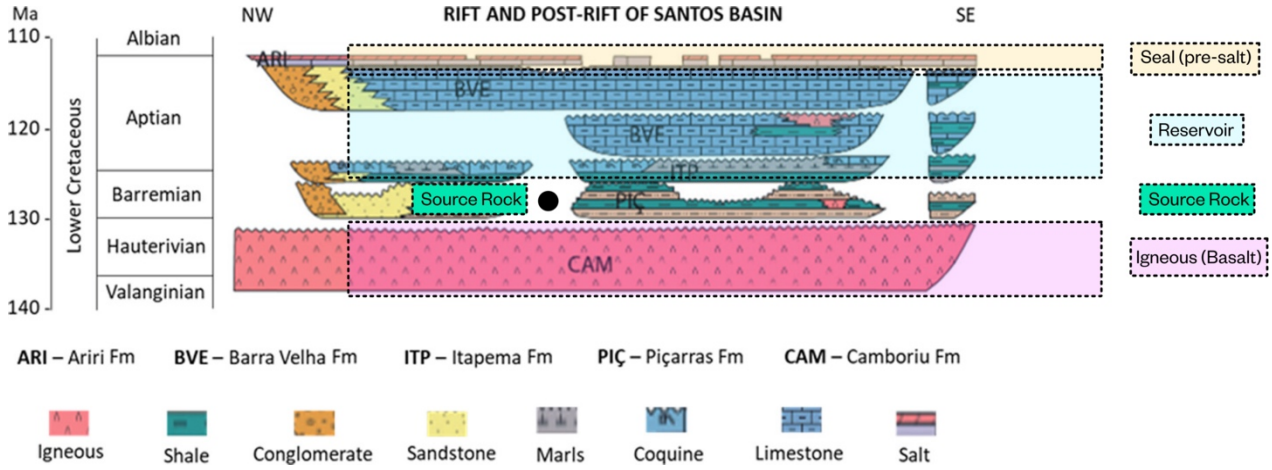

Figure 2-6: Stratigraphic chart of the rift and post-rift phases of the Santos Basin, modified (Link)

2.7 Locations and Well-log plots of different wells:

The wells 3-BRSA-944A-RJS, 1-BRSA-1116-RJS, 3-BRSA-1215-RJS, and 9-ITP-1-RJS are strategically located in the vicinity of Rio de Janeiro, constituting the Itapù field, with the exception of 3-BRSA-944A-RJS, which is affiliated with the Búzios field. These wells play a pivotal role in the exploration and production activities within the oil and gas industry, focusing primarily on the Barra Velha Formation. This formation, recognized as a pre-salt carbonate reservoir, encompasses limestones and microbialites, contributing significantly to the energy resources in the region (figures 2-7, 2-8).

Specifically, 1-BRSA-1116-RJS represents an exploration well, designed to identify and characterize new reservoirs of oil or gas. Meanwhile, 3-BRSA-1215-RJS serves as an appraisal well, instrumental in evaluating the size, production rate, and reserves of the oil or gas field. The production well, 9-ITP-1-RJS, is crucial for the extraction of oil or gas from the Barra Velha Formation, ensuring a sustainable supply of energy resources.

On the other hand, the wells 3-EQNR-1-SPS and 3-EQNR-3-SPS are strategically positioned near Sao Paulo, contributing to the Bacalhau field, formerly known as Carcara. Similar to their Rio de Janeiro counterparts, these wells tap into the Barra Velha Formation. Notably, they exhibit a higher gas-oil ratio compared to other wells in the region, introducing a distinctive element to the exploration and production dynamics.

In this context, the abbreviations are integral to understanding the geographical and operational aspects:

- BRSA signifies Brazil
- RJS denotes the state of Rio de Janeiro in Brazil
- SPS represents the state of Sao Paulo in Brazil

This exploration and production framework, characterized by exploration, appraisal, and production wells, underscores the significance of these wells in advancing the oil and gas industry in the region.

The unique geological characteristics of the Barra Velha Formation further emphasize the importance of these assets in securing and optimizing energy resources for sustainable development.

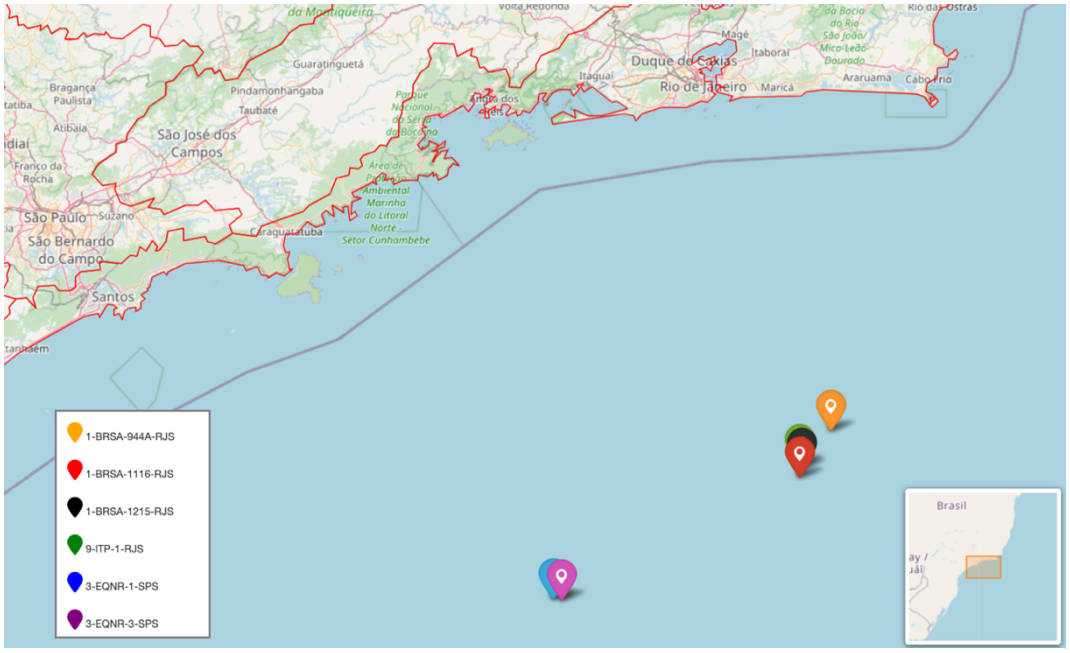

Figure 2-7: Map of six different wells [3-BRSA-944A-RJS, 1-BRSA-1116-RJS, 3-BRSA-1215-RJS, and 9-ITP-1-RJS, 3-EQNR-1-SPS and 3-EQNR-3-SPS] in fields of Itapù, Bùzios, Bacalhau (all located in the Santos Basin)

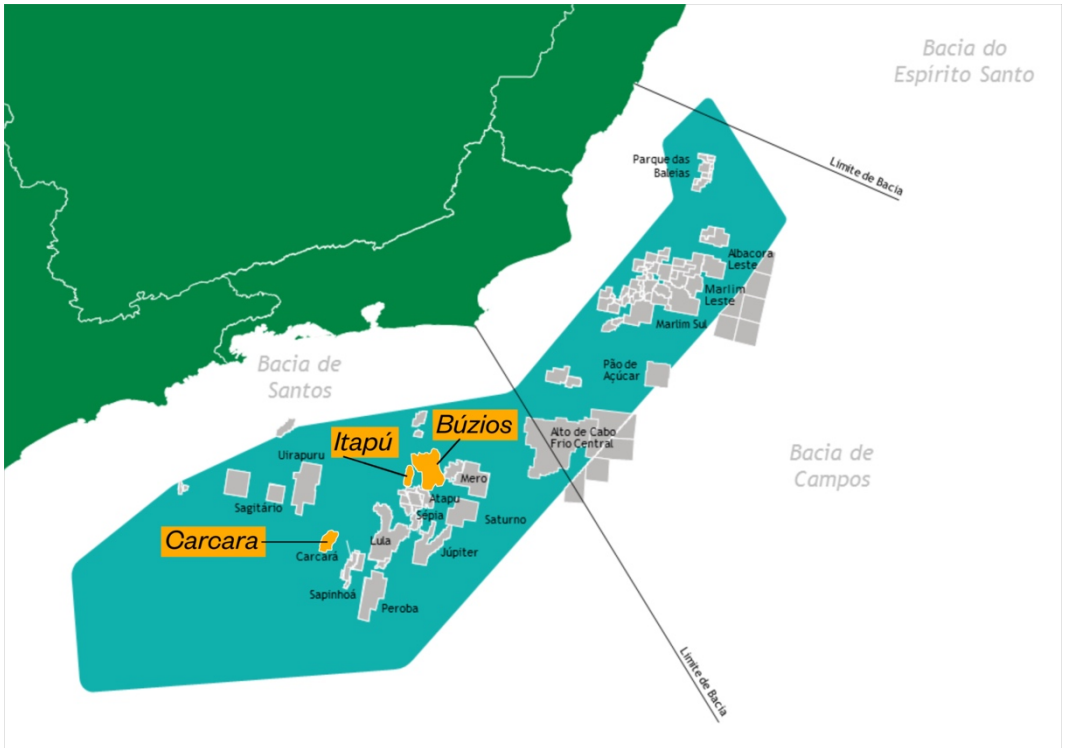

Figure 2-8: Map of three different fields (Itapù, Bùzios, Bacalhau); Six different wells are drilled into these three fields.

The prediction of permeability poses a significant challenge, especially in carbonate reservoirs, which typically exhibit lower levels of permeability compared to their sandstone counterparts. The complexity is further compounded by the heterogeneous distribution of vugs and fractures, which significantly contribute to the overall permeability of carbonate reservoirs. Carbonate reservoirs are composed of rocks that contain calcium carbonate minerals, such as calcite and dolomite. They are often more complex and heterogeneous than sandstone reservoirs, which consist of quartz and other silicate minerals. You can observe the pore size and pore distribution in carbonate reservoirs in the Santos Basin, specifically for well 9-ITP-RJS, in the provided image (figure 2-9).

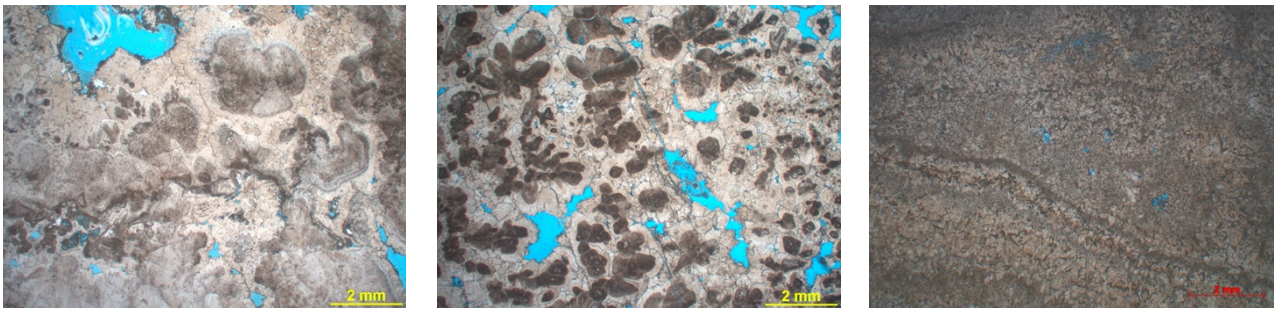

*Figure 2-9: **Pore Distribution in Well 9-ITP-RJS, Santos Basin: A close-up view showcasing the intricate pore structure and distribution.***

In Well 3-BRSA-944A-RJS, the geological formations encountered mirrored those in Well 1-BRSA-1116-RJS to a significant extent. The Ariri formation, between 3444 to 6003.6 meters, shared similarities with the corresponding depths in the first well, displaying high salinity, lack of porosity, and limited reservoir potential due to its composition of anhydrite and halite. The Barra Velha formation in this well, ranging from 5487 to 5782 meters, resembled the favorable reservoir characteristics found in Well 1, showcasing limestone lithology with high porosity and permeability. Subsequently, the Itapema formation, occupying 5782 to 6003.6 meters, echoed the same limitations observed in the corresponding depths of Well 1, featuring calcareous shale with insufficient porosity and permeability for effective fluid flow (figure 2-10).

In Well 1-BRSA-1116-RJS, the formations encountered spanned varied lithologies and reservoir potentials. The Ariri formation, ranging from 4050 to 5940 meters, consisted mainly of anhydrite and halite. Despite its substantial thickness of 1292 meters, this formation lacked porosity, rendering it an inadequate reservoir rock due to its high salinity and low fluid-transmitting capabilities. Following the Ariri, the Barra Velha formation, occupying 5342 to 5857 meters, presented limestone lithology spanning 515 meters. Contrarily, this formation exhibited high porosity and permeability, making it an excellent reservoir rock for fluid storage and transmission. Lastly, the Itapema formation, lying between 5857 to 5940 meters, predominantly comprised calcareous shale, characterized by low porosity and permeability, making it unsuitable for fluid flow (figure 2-11).

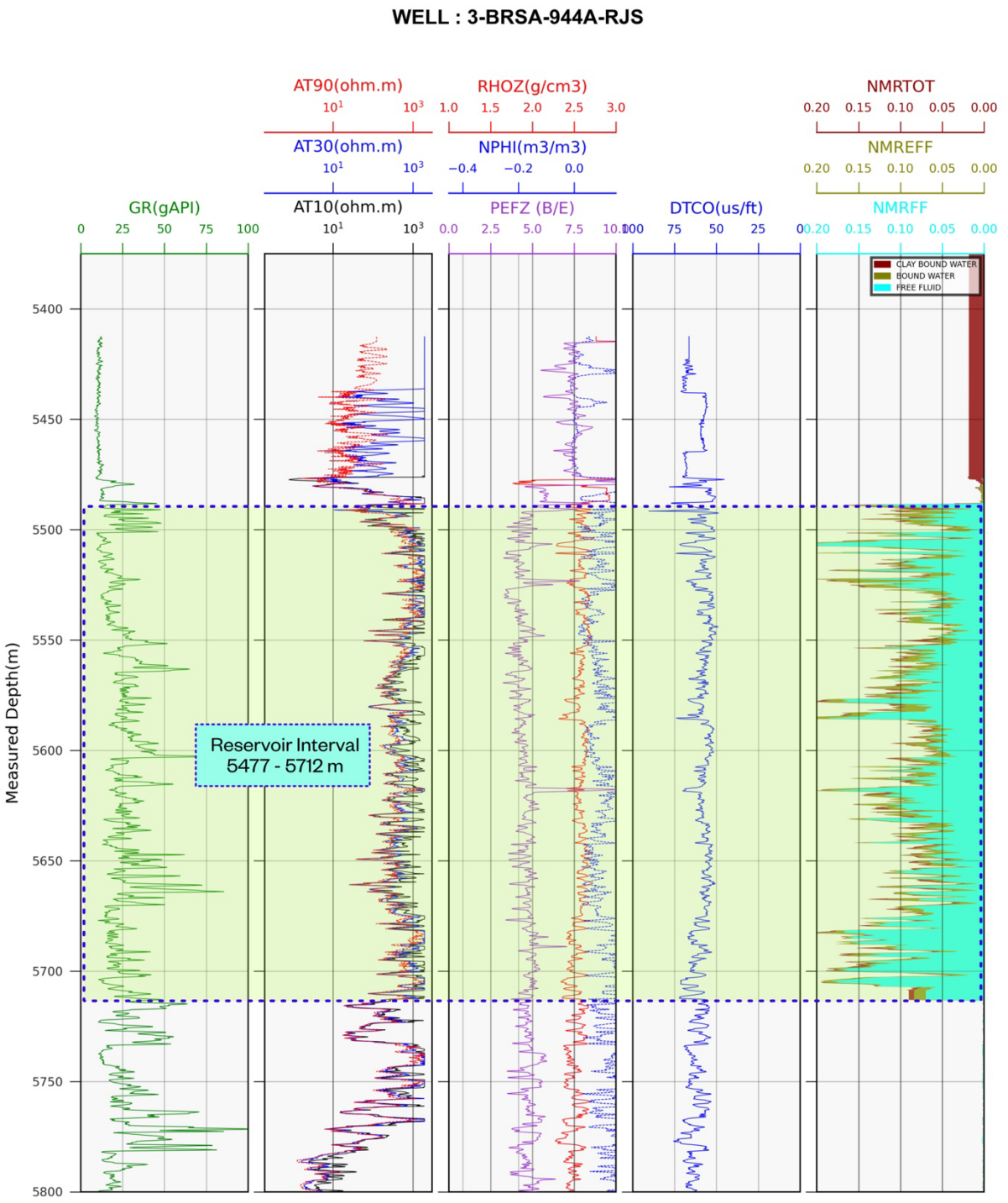

Figure 2-10: Composite well-log of well 3-BRSA-944A-RJS (prediction depth interval: 5477 - 5712 m) vs Measured Depth. Track 1: Gamma-Ray (GR). Track 2: Induction Electric Resistivity logs. Investigation depths of 10 (AT10), 30 (AT30), and 90 (AT90) inches. Track 3: Formation Density (RHOZ), Neutron Porosity (NPHI), and Photoelectric Factor (PEFZ) logs. Track 4: Compressional Wave Slowness (DTCO). Track 5: Nuclear Magnetic Resonance Porosity logs. Total Porosity (NMRTOT), Effective Porosity (NMREFF), and Free Fluid (NMRFF). The light green highlight interval which is the reservoir rock.

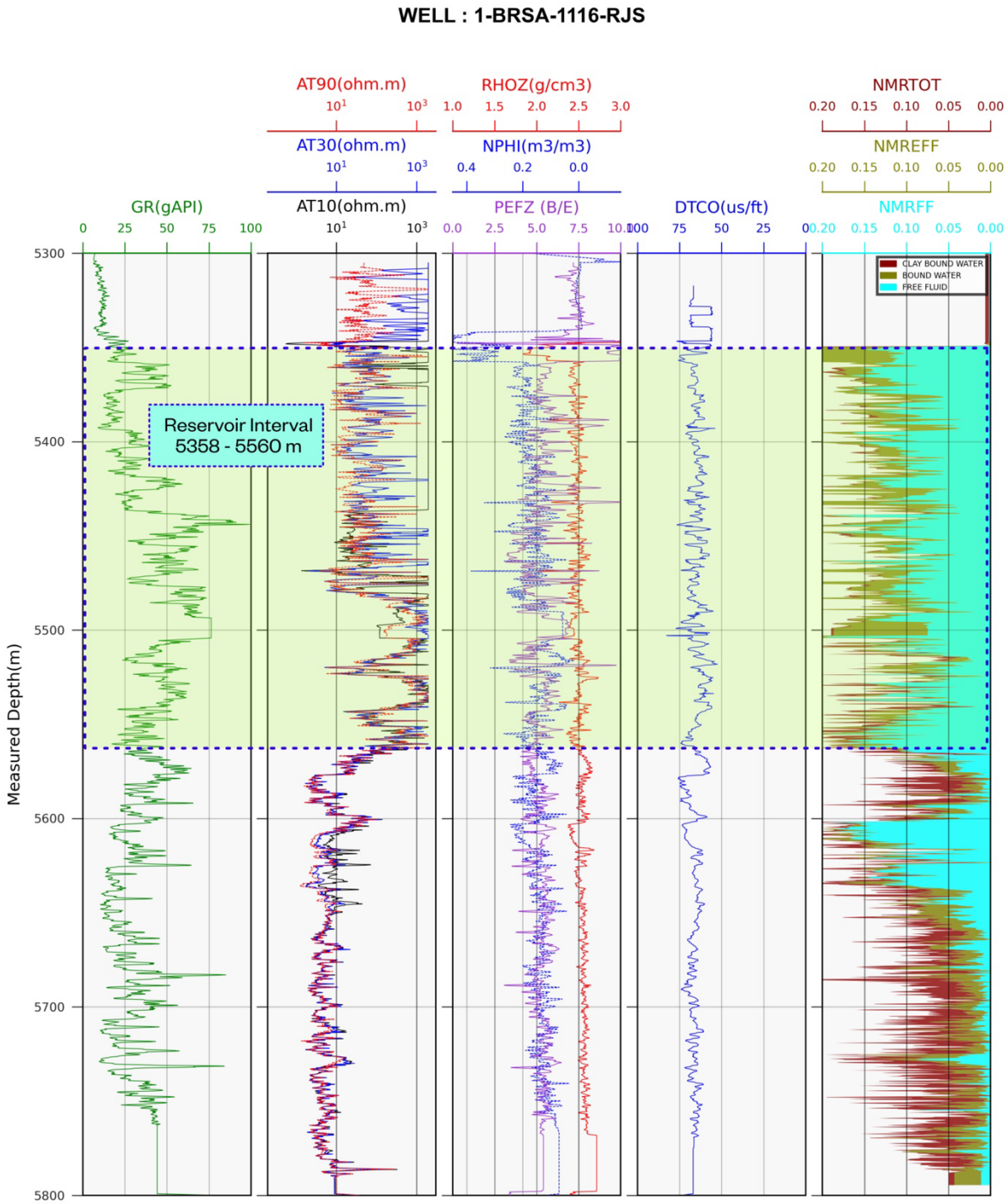

Figure 2-11: Composite well-log of well 1-BRSA-1116-RJS (prediction depth interval: 5358 - 5560 m) vs Measured Depth. Track 1: Gamma-Ray (GR). Track 2: Induction Electric Resistivity logs. Investigation depths of 10 (AT10), 30 (AT30), and 90 (AT90) inches. Track 3: Formation Density (RHOZ), Neutron Porosity (NPHI), and Photoelectric Factor (PEFZ) logs. Track 4: Compressional Wave Slowness (DTCO). Track 5: Nuclear Magnetic Resonance Porosity logs. Total Porosity (NMRTOT), Effective Porosity (NMREFF), and Free Fluid (NMRFF). The light green highlight interval which is the reservoir rock.

In Well 3-BRSA-1215-RJS, the geological profile followed a pattern reminiscent of the previous wells. The Ariri formation, spanning 4709 to 5681 meters, exhibited similarities with Wells 1 and 3 in terms of high salinity, lack of porosity, and unsuitability as a reservoir rock due to its anhydrite and halite

content. The subsequent Barra Velha formation, occupying 5412 to 5681 meters, mirrored the favorable characteristics observed in Wells 1 and 3, presenting limestone lithology with notable porosity and permeability, rendering it a favorable reservoir rock for fluid storage and transmission (figure 2-12).

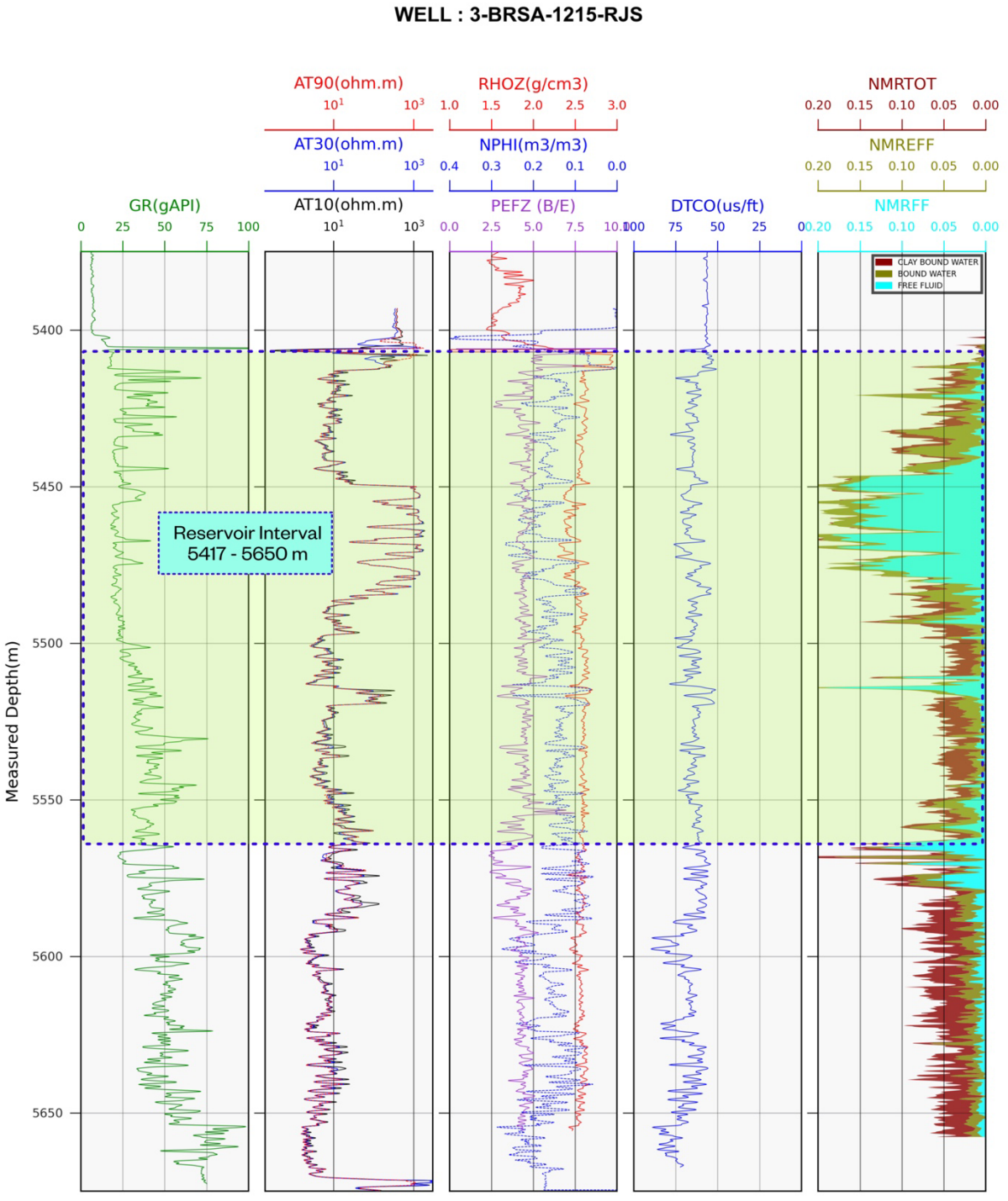

Figure 2-12: Composite well-log of well 3-BRSA-1215-RJS (prediction depth interval: 5417 - 5650 m) vs Measured Depth. Track 1: Gamma-Ray (GR). Track 2: Induction Electric Resistivity logs. Investigation depths of 10 (AT10), 30 (AT30), and 90 (AT90) inches. Track 3: Formation Density (RHOZ), Neutron Porosity (NPHI), and Photoelectric Factor (PEFZ) logs. Track 4: Compressional Wave Slowness (DTCO). Track 5: Nuclear Magnetic Resonance Porosity logs. Total Porosity (NMRTOT), Effective Porosity (NMREFF), and Free Fluid (NMRFF). The light green highlight interval which is the reservoir rock.

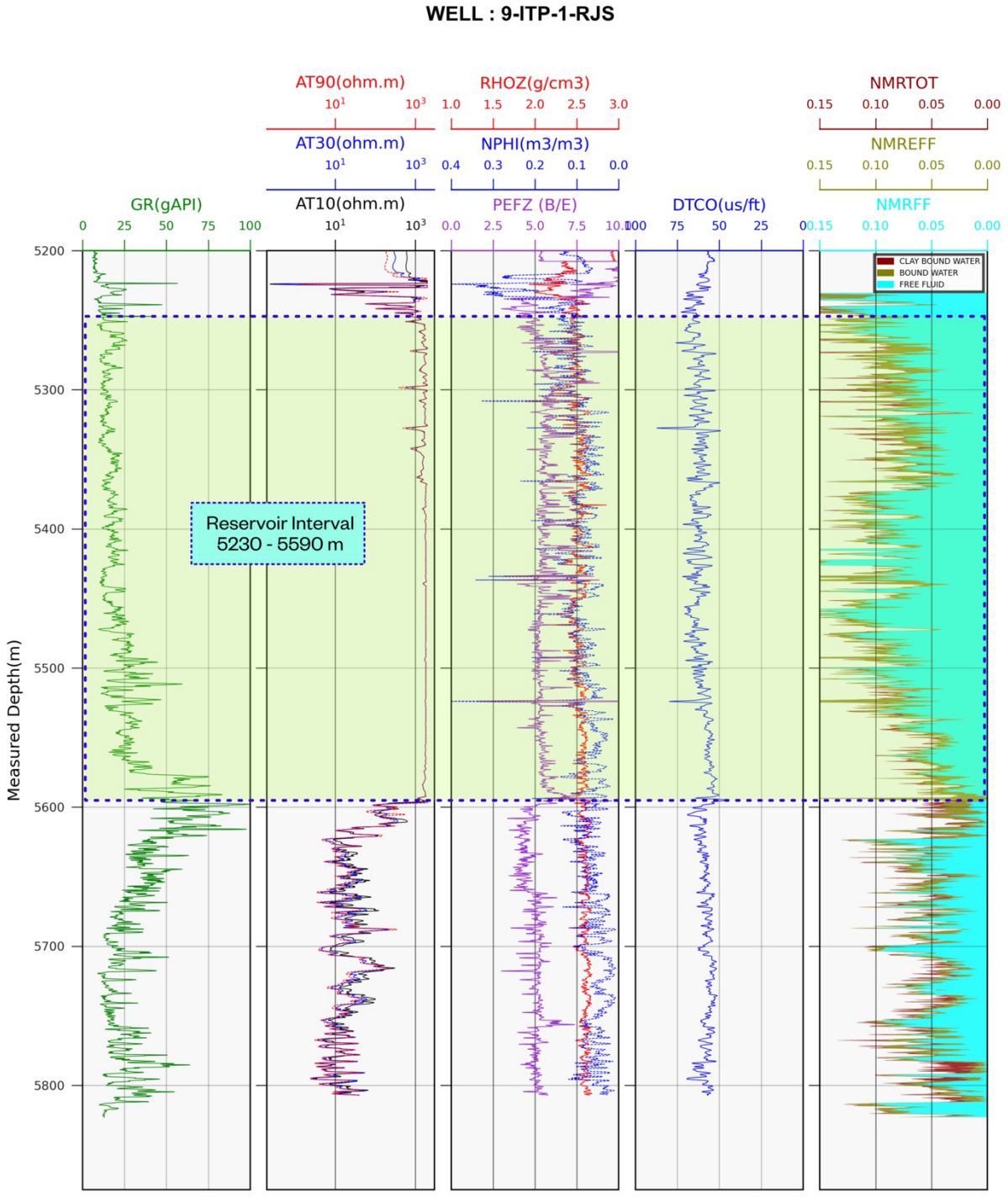

Figure 2-13: Composite well-log of well 9-ITP-1-RJS (prediction depth interval: 5230 - 5590 m) vs Measured Depth. Track 1: Gamma-Ray (GR). Track 2: Induction Electric Resistivity logs. Investigation depths of 10 (AT10), 30 (AT30), and 90 (AT90) inches. Track 3: Formation Density (RHOZ), Neutron Porosity (NPHI), and Photoelectric Factor (PEFZ) logs. Track 4: Compressional Wave Slowness (DTCO). Track 5: Nuclear Magnetic Resonance Porosity logs. Total Porosity (NMRTOT), Effective Porosity (NMREFF), and Free Fluid (NMRFF). The light green highlight interval which is the reservoir rock.

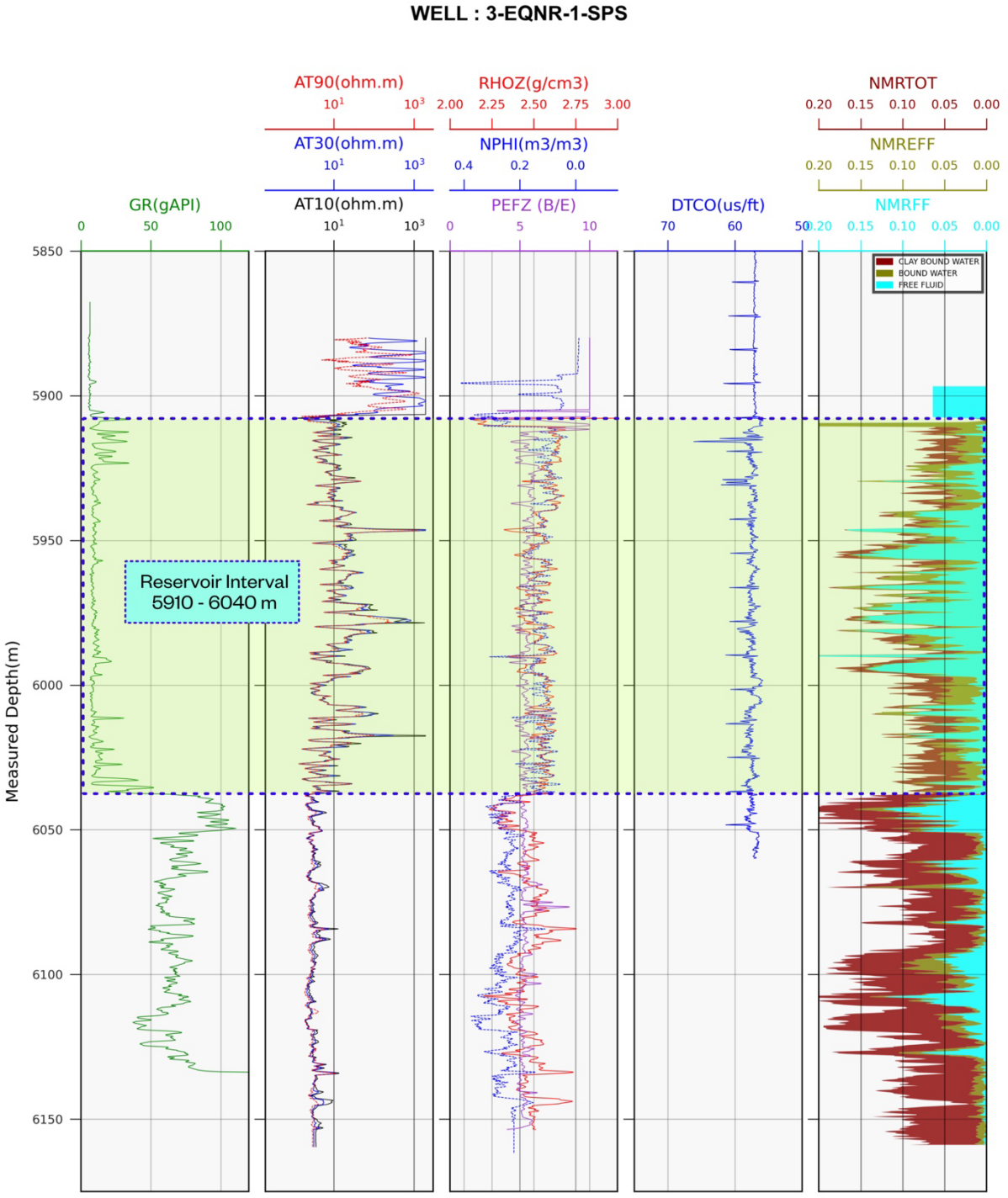

Figure 2-14: Composite well-log of well 3-EQNR-1-SPS (prediction depth interval: 5910 - 6040 m) vs Measured Depth. Track 1: Gamma-Ray (GR). Track 2: Induction Electric Resistivity logs. Investigation depths of 10 (AT10), 30 (AT30), and 90 (AT90) inches. Track 3: Formation Density (RHOZ), Neutron Porosity (NPHI), and Photoelectric Factor (PEFZ) logs. Track 4: Compressional Wave Slowness (DTCO). Track 5: Nuclear Magnetic Resonance Porosity logs. Total Porosity (NMRTOT), Effective Porosity (NMREFF), and Free Fluid (NMRFF). The light green highlight interval which is the reservoir rock.

Well 3-EQNR-3-SPS revealed a significant geological shift beyond 6315 meters, marked by elevated gamma ray log values and decreased resistivity log values, attributed to the presence of igneous rock, specifically basalt formations. Basalt's mineral composition contributes to higher gamma radiation

readings, while its inherent lower resistivity, compared to sedimentary reservoir rocks, is reflected in the decreased resistivity log values. Despite these log indications, basalt formations typically lack the necessary porosity and permeability for effective reservoir capabilities, thereby rendering them unsuitable as reservoir rocks (figure 2-14).

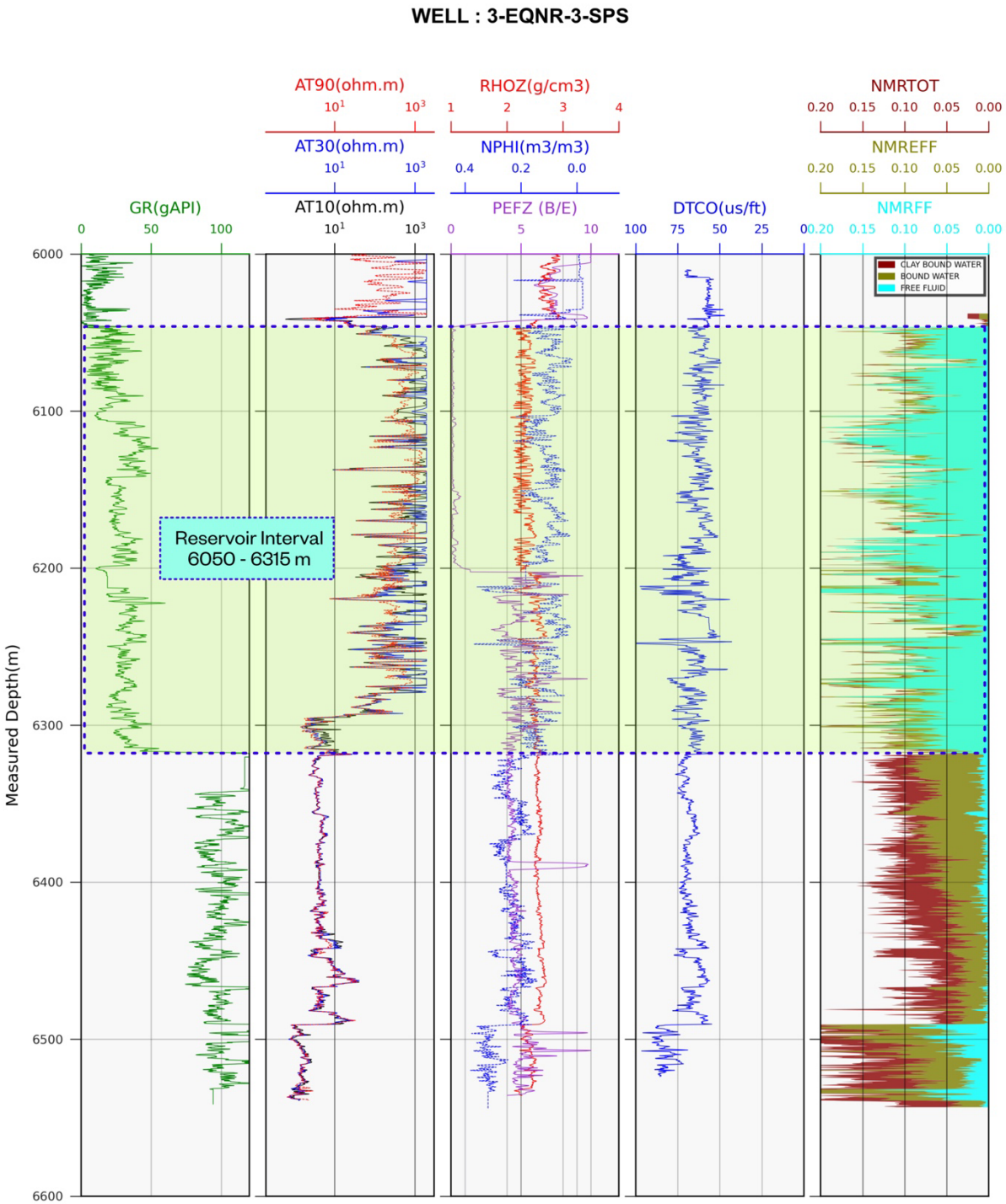

Figure 2-15: Composite well-log of well 3-EQNR-3-SPS (prediction depth interval: 6050 - 6315 m) vs Measured Depth. Track 1: Gamma-Ray (GR). Track 2: Induction Electric Resistivity logs. Investigation depths of 10 (AT10), 30 (AT30), and 90 (AT90) inches. Track 3: Formation Density (RHOZ), Neutron Porosity (NPHI), and Photoelectric Factor (PEFZ) logs. Track 4: Compressional Wave Slowness (DTCO). Track 5: Nuclear Magnetic Resonance Porosity logs. Total Porosity (NMRTOT), Effective Porosity (NMREFF), and Free Fluid (NMRFF). The light green highlight interval which is the reservoir rock.

2.8 Visualizing Well Log Relationships and Distributions

This section investigates the relationships and distributions of several key well log parameters that are commonly used in the oil and gas industry to evaluate the subsurface formations and their fluid contents. These parameters are: gamma ray (GR), deep resistivity (AT90), bulk density (RHOZ), neutron porosity (NPHI), compressional wave slowness (DTCO), photoelectric factor (PEFZ), free fluid nuclear magnetic resonance (FFNMR), effective porosity nuclear magnetic resonance (EFFNMR), and total porosity nuclear magnetic resonance (TOTNMR).

The histograms provide insights into the distributional characteristics of each parameter, including the central tendency, spread, and shape of the data. For instance, a bimodal histogram of GR may indicate the presence of shale and sand layers, while a skewed histogram of AT90 may reflect the variation of resistivity due to different fluid saturations. This analysis provides valuable insights into the behavior of these parameters and their potential applications in reservoir characterization and fluid property estimation. By using these graphical tools, it can be identified the lithology, porosity, permeability, and hydrocarbon potential of the reservoir, as well as detect any anomalies or outliers in the data (figures 2-16, 2-17).

A comprehensive visual analysis is conducted using Seaborn pairplots and histograms to unveil the interdependencies and patterns of these parameters. The pairplots show the scatter plots of each pair of parameters, as well as the histograms of each parameter along the diagonal. The pairplots reveal the correlations between the log parameters, highlighting the strength and direction of these relationships. For example, a positive correlation between RHOZ and PEFZ indicates a higher density and photoelectric effect of the rock matrix, while a negative correlation between NPHI and DTCO suggests a lower porosity and faster acoustic velocity of the formation (figures 2-18, and 2-19).

Also, the importance of ensuring that each well-log data falls within the same range, rather than having different scales for all six wells, becomes apparent. For instance, having varying gamma-ray (GR) log ranges for each of the six wells indicates the presence of distinct geological models and petrophysical regimes in a real-world scenario. This variation can potentially highlight the unreliability of the final model. Therefore, different histograms, kernel density estimates (KDE), and strip plots were utilized to demonstrate the consistent range for the specific well-log data across all six wells.

In the process of developing the model, the challenge posed by the existence of diverse well-log data for each well was addressed through Min-Max normalization. This technique effectively brought all the well-log data into the same range, ensuring uniformity and comparability across different datasets.

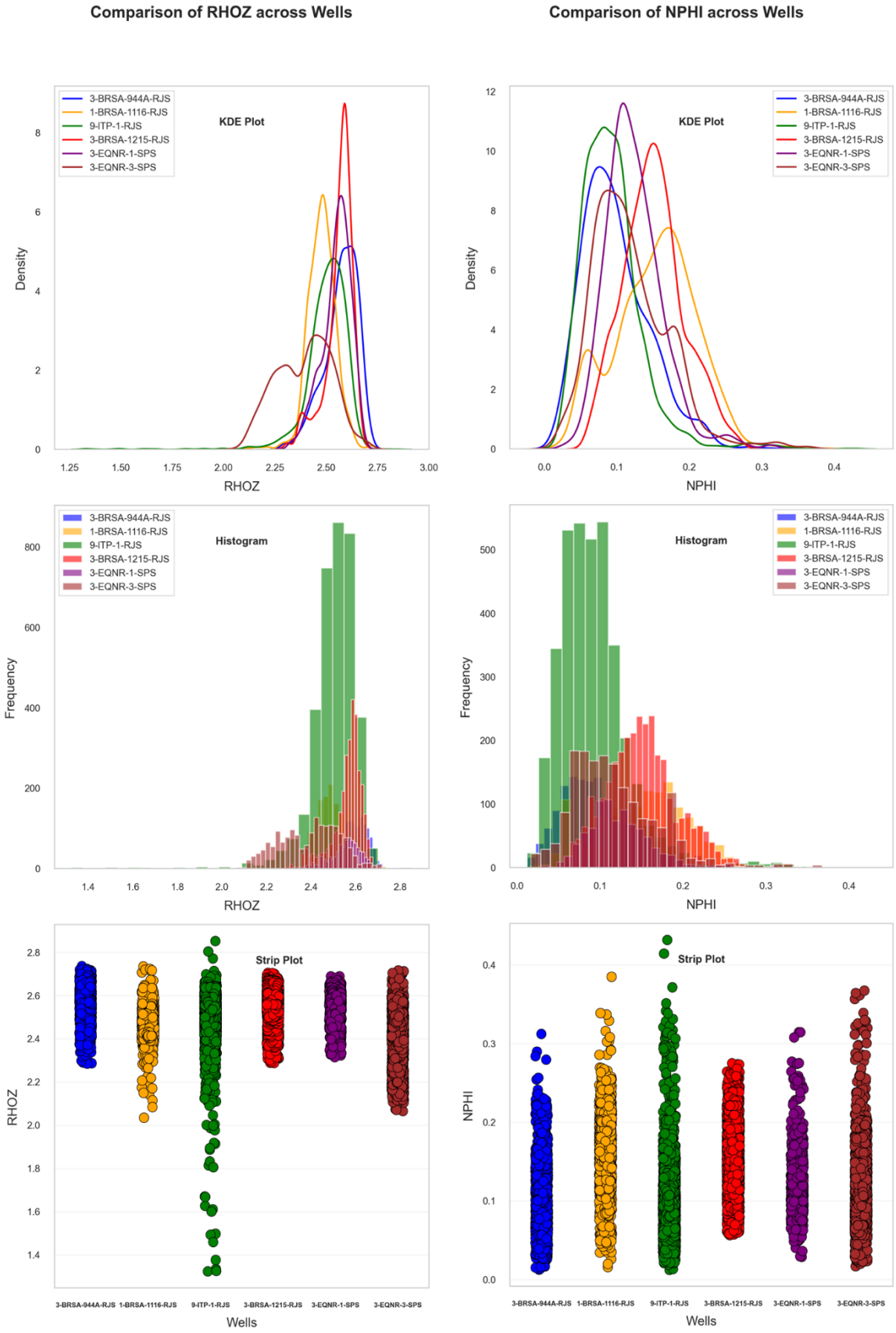

Figure 2-16: Different Plots (KDE Plot, Histogram, and Strip) for RHOZ on the left and NPHI on the right side. Each of these plots for different properties (RHOZ & NPHI) show the distribution and comparison the range of the parameters for six different wells.

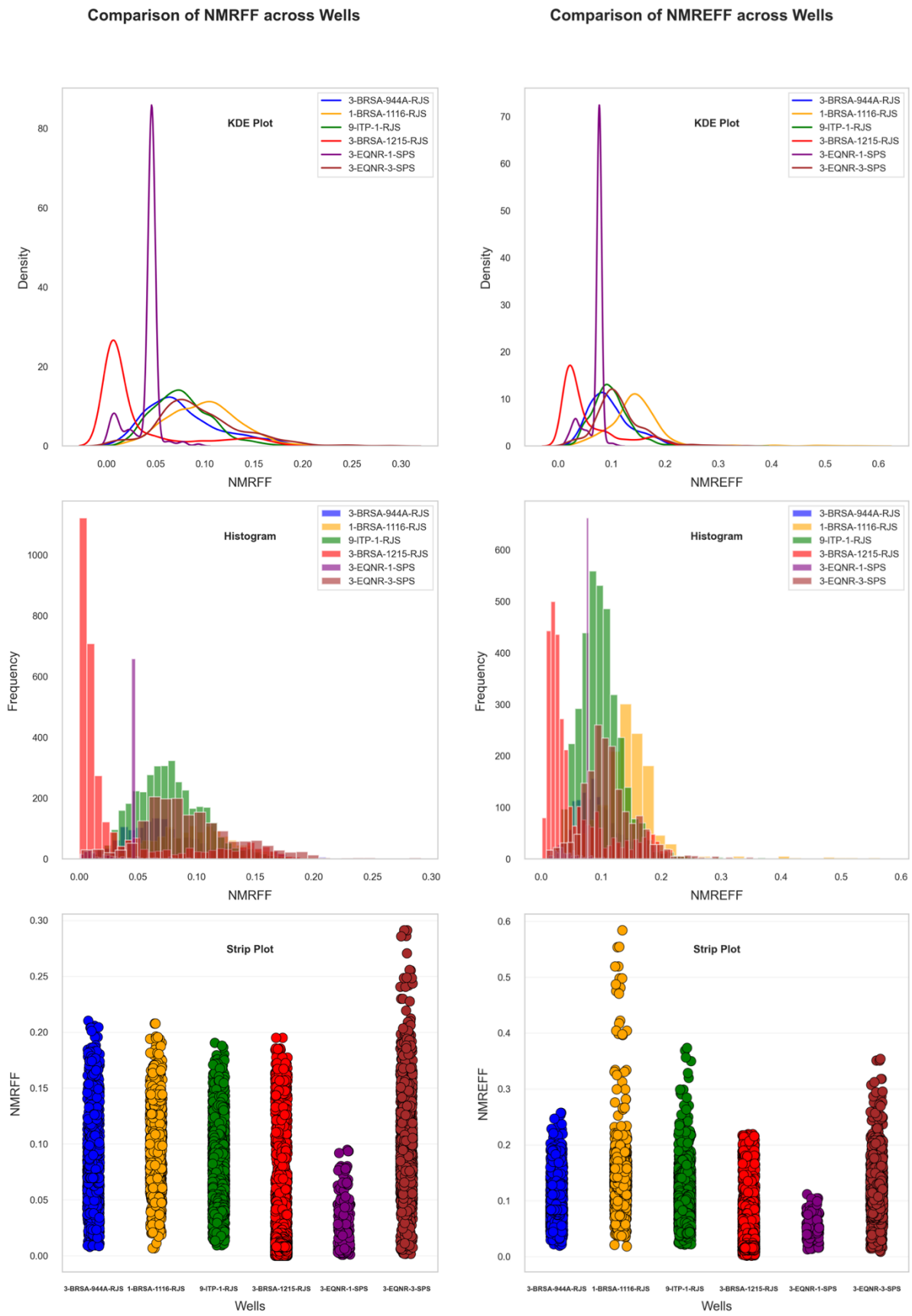

Figure 2-17: Different Plots (KDE Plot, Histogram, and Strip) for NMRFF on the left and NMREFF on the right side. Each of these plots for different properties (NMRFF & NMREFF) show the distribution and comparison the range of the parameters for six different wells.

2.8.1 Part One Correlations

In this section, the correlation between various basic well-log data and NMR log data is depicted using Seaborn Pair plots. This directly impacts the quality of the predictive model we aim to develop. The correlation can be positive or negative; a positive correlation indicates a meaningful relationship between the logs to some extent.

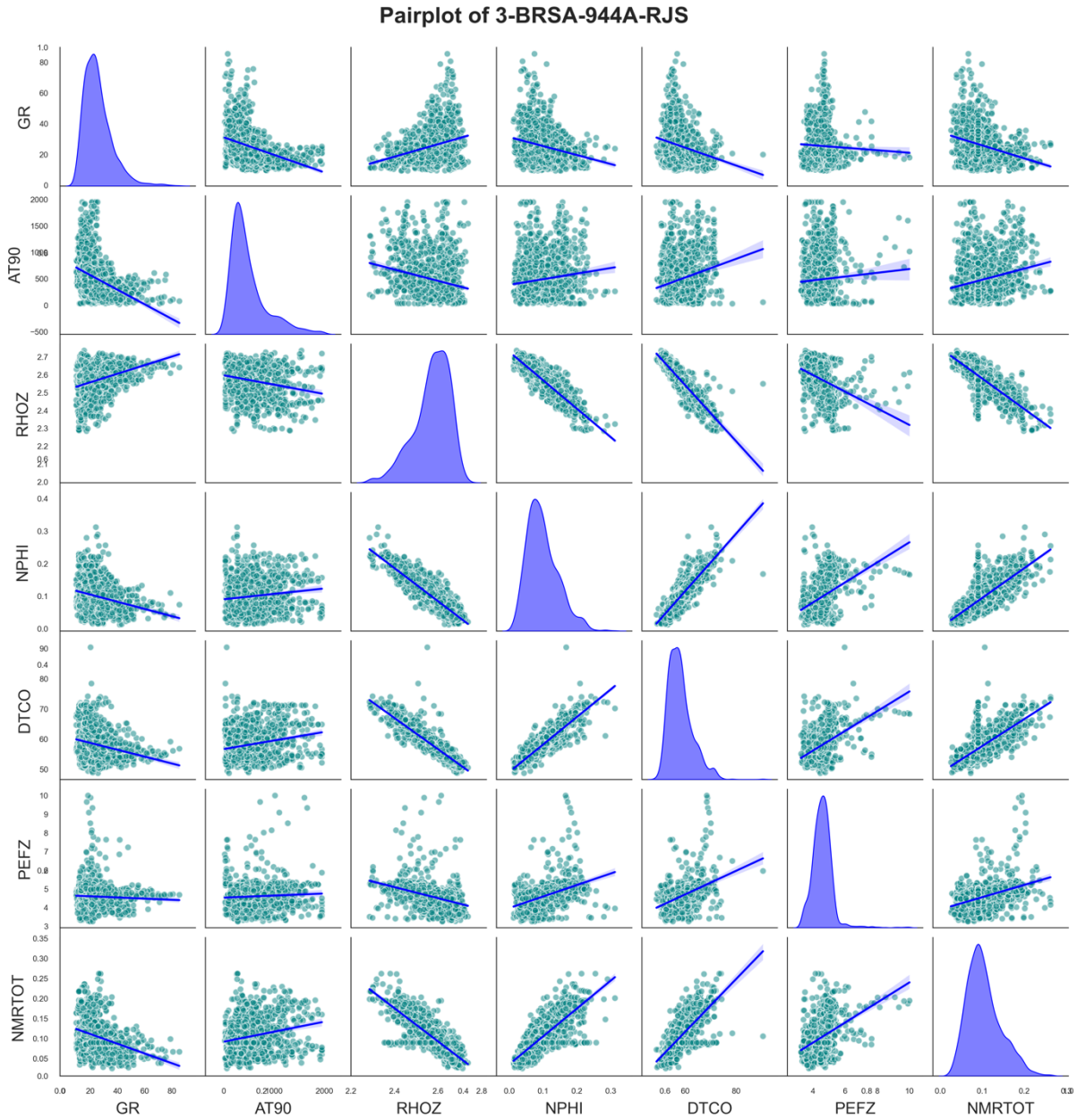

Figure 2-18: Seaborn Pair plots of well: 3-BRSA-944A-RJS for different well log data (GR, AT90, RHOZ, NPHI, DTCO, PEFZ, and NMRTOT)

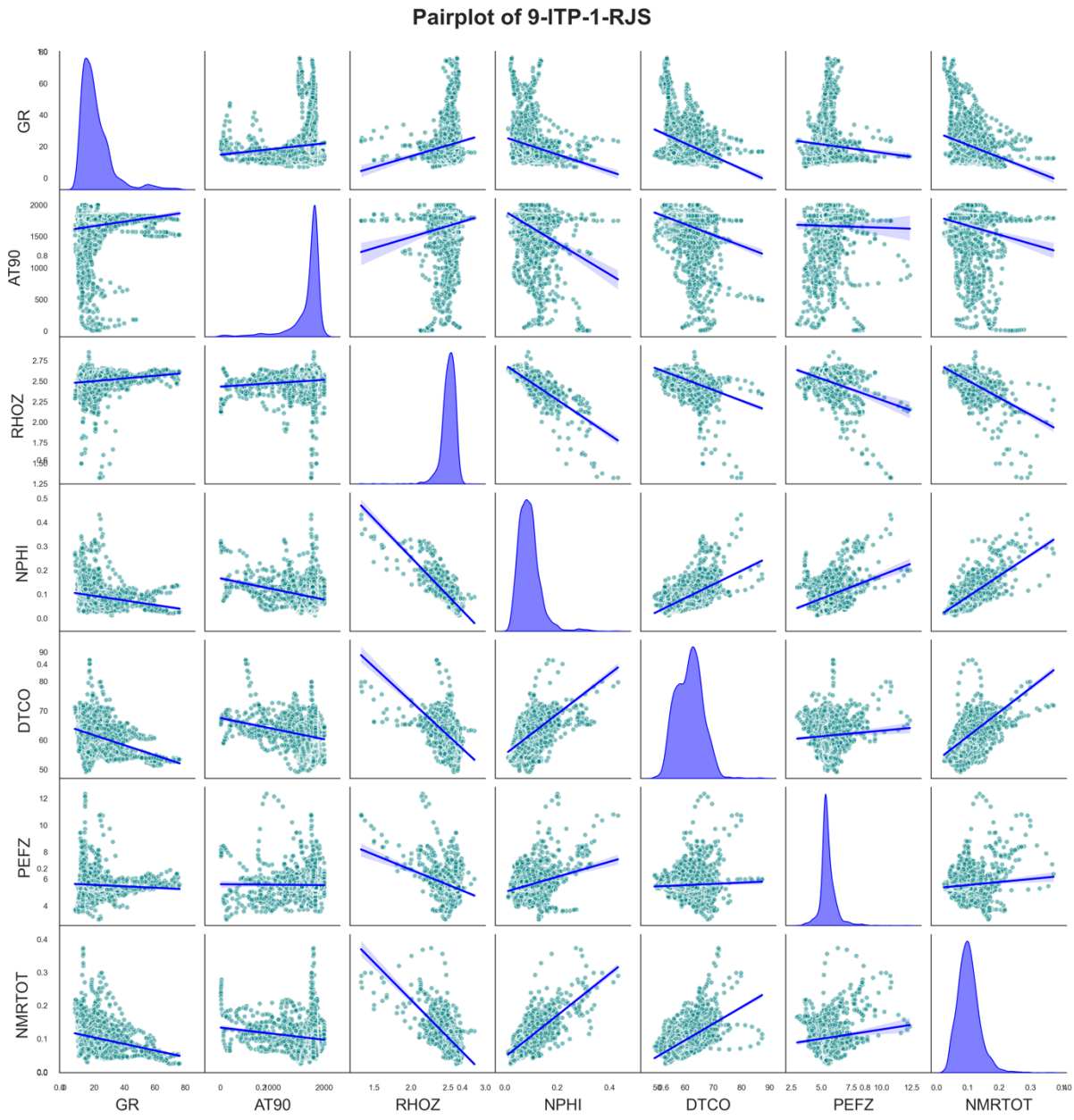

Figure 2-19: Seaborn Pair plots of well: 9-ITP-1-RJS for different well log data (GR, AT90, RHOZ, NPHI, DTCO, PEFZ, and NMRTOT)

Other Histograms and Distributions plot have been shown in Appendix 01.

2.8.2 Part Two Correlations

In this section, we have explored the relationship between fundamental well-log data and permeability. The analysis delves into understanding how various essential well-log measurements correspond or relate to the permeability of the formation. By examining this correlation, we aim to uncover the extent to which these basic well-log measurements serve as indicators or predictors of permeability within the geological formations under study (figure 2-20).

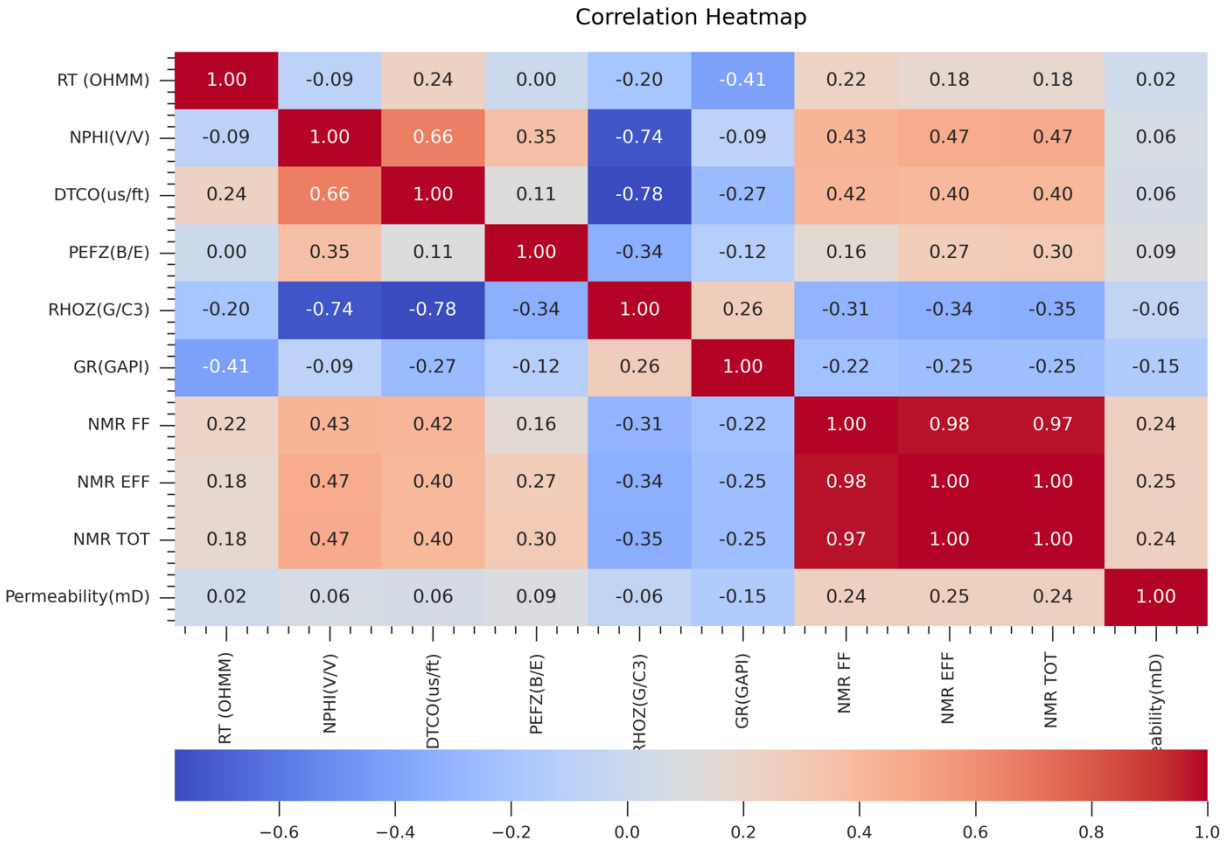

Figure 2-20: Heatmap of correlation of well-log data and laboratory permeability related to the second part of the thesis

3 Methodologies and Concepts

3.1 Comprehensive Approaches to Permeability Estimation: Leveraging NMR and Well-Log Data Analysis in Predictive Modeling

In the first part of the study, the permeability estimation process takes an indirect approach by using Nuclear Magnetic Resonance (NMR) data. The predictors are based on fundamental well-log data, while the target for estimation is the NMR-derived permeability. By employing deterministic correlation, such as the Timur-Coates model (Jun Yoneda, 2022), the prediction of permeability is enabled based on the anticipated NMR values. This approach is rooted in the understanding that NMR data can offer valuable insights into the pore structure and fluid properties of subsurface formations, which are crucial for estimating permeability.

Using NMR data in permeability estimation offers several advantages. NMR provides direct measurements of fluid-filled porosity and pore size distribution, which are crucial factors influencing permeability. By incorporating NMR data into the predictive models, the estimation of permeability can benefit from a more comprehensive understanding of the reservoir properties, leading to more accurate predictions. Additionally, NMR data can capture variations in fluid types and their distribution within the reservoir, contributing to a more nuanced assessment of permeability across different geological settings. This means that by using NMR data, we can get a better understanding of the reservoir and make more accurate predictions of permeability, which is essential for various applications in the oil and gas industry.

In Part Two of the study, a different approach is taken by excluding NMR data from the well-log predictors. Instead of relying on NMR-derived values, the focus shifts to predicting permeability directly using laboratory-derived permeability as the target variable. This unique strategy allows for a distinct examination of the predictive models and methodologies when NMR-derived data is not included in the prediction process, offering insights into the direct estimation of permeability from well-log data without relying on the intermediary step of NMR-derived predictions. This approach provides a comprehensive analysis of the effectiveness of well-log data alone in predicting permeability, highlighting the strengths and limitations of using NMR-derived data versus relying solely on well-log data for permeability estimation.

The exclusion of NMR data from the well-log predictors in Part Two presents an opportunity to assess the effectiveness of well-log data alone in predicting permeability. This approach is significant as it provides a comparative analysis of the predictive models and methodologies, highlighting the strengths and limitations of using NMR-derived data versus relying solely on well-log data for permeability estimation.

The two parts of the study offer a comprehensive analysis of permeability estimation using different approaches. Part One demonstrates the indirect estimation of permeability through NMR-derived data, while Part Two focuses on the direct estimation of permeability from well-log data without the intermediary step of NMR-derived predictions. This comparative analysis provides valuable insights

into the role of NMR data in permeability estimation and the potential of well-log data as a standalone predictor for permeability.

The foundation for predicting permeability based on well-log data involves a comprehensive understanding of essential concepts such as the utilization of Python libraries and the application of various model development and optimization techniques. These preliminary discussions serve as crucial building blocks, providing the necessary groundwork before delving into the intricate process of permeability prediction.

Python libraries play a pivotal role in providing a versatile toolkit for data manipulation, statistical analysis, and model implementation. Libraries such as NumPy, Pandas, and Scikit-learn offer powerful tools for handling and processing large datasets, performing statistical analysis, and implementing machine learning algorithms for predictive modeling. The utilization of these libraries enables researchers and practitioners to efficiently preprocess well-log data, extract relevant features, and build predictive models for permeability estimation.

Furthermore, the application of various model development and optimization techniques is essential for ensuring the accuracy and efficiency of the predictive models. Techniques such as feature engineering, model selection, hyperparameter tuning, and cross-validation are integral components of the model development process. Feature engineering involves transforming raw well-log data into meaningful predictors that capture the underlying geological properties influencing permeability. Model selection entails choosing the most appropriate algorithm for the predictive task, considering factors such as interpretability, scalability, and predictive performance. Hyperparameter tuning aims to optimize the configuration of the selected model to achieve the best predictive accuracy. Cross-validation techniques are employed to assess the generalization performance of the models and mitigate overfitting.

3.2 Python library

The thesis effectively utilized a suite of Python libraries to address the challenges of data handling, visualization, and machine learning in scientific research. The following Python libraries were employed for different tasks:

Data Manipulation and Analysis:

pandas (pd): This extensive library served as the backbone for handling the massive volume of well log data. Its DataFrames data structure provided a powerful and flexible framework for organizing, filtering, and manipulating the data, enabling efficient data exploration and analysis.

numpy (np): This fundamental library provided essential numerical capabilities for performing array operations, calculating statistics, and applying mathematical functions on the extracted numerical values from the well log data. Its efficient and versatile algorithms significantly streamlined data processing tasks.

Plotting and Visualization:

matplotlib.pyplot (plt): This foundational module within Matplotlib played a central role in creating a wide range of visualizations to effectively communicate insights from the data. Its comprehensive plotting capabilities enabled the creation of insightful charts, graphs, and plots, facilitating data exploration and interpretation.

seaborn (sns): Building upon Matplotlib's strengths, seaborn offered advanced plotting features, empowering the creation of aesthetically pleasing and informative statistical graphics. Its ability to handle categorical data, create joint plots, and provide visual themes enhanced the overall presentation and communication of data-driven findings.

matplotlib.patches (mpatches): This module within Matplotlib was employed to further enhance the visual appeal of plots, particularly through the creation of colored boxes for legends.

Customization and Formatting:

matplotlib_inline.backend_inline.set_matplotlib_formats: This configuration ensured that Matplotlib figures were displayed seamlessly within the Jupyter Notebook environment as scalable vector graphics (SVG). This enabled high-quality image rendering without compromising readability.

qbstyles.mpl_style: This potentially custom style module was applied to Matplotlib plots, customizing their appearance and presentation to align with the project's visual style guidelines.

Machine Learning and Preprocessing:

scikit-learn (sklearn): This extensive machine learning library provided the necessary tools for building and evaluating regression models to predict target values from the well log data. Its comprehensive range of regression algorithms, including RandomForestRegressor and GradientBoostingRegressor, enabled the exploration of various modeling approaches.

Pickle: This module enabled the serialization and deserialization of Python objects, facilitating the efficient saving and loading of trained models, preprocessed data, and other intermediate results. Its

ability to handle large datasets and preserve data integrity was crucial for maintaining reproducible research workflows.

The use of these libraries facilitated efficient exploration and analysis of well log data, creation of insightful visualizations, and development of robust machine learning models for predicting target values from the data.

3.3 Model Development and Optimization Techniques

Data preparation plays a critical role in machine learning by ensuring the quality and consistency of the data. It involves identifying and addressing data issues such as missing values, outliers, and data inconsistencies. By ensuring the data is clean and well-structured, the machine learning algorithm can learn more effectively and produce more accurate predictions.

3.3.1 Normalization

Normalization techniques like Min-Max scaling and StandardScaler are crucial for ensuring robustness in machine learning models by standardizing variable scales (Ahmed M. Elshewey, 2023). These methods address challenges posed by predictors with varying magnitudes, enabling effective learning across diverse feature distributions.

Min-Max scaling, or z-score normalization, adjusts data points to a mean of 0 and a standard deviation of 1. This compression of variable ranges facilitates a consistent understanding of feature relationships, allowing the model to focus on relative differences between features rather than individual magnitudes.

StandardScaler, in contrast, standardizes features by removing the mean and scaling to unit variance. Unlike Min-Max scaling, it doesn't confine values to a specific range but ensures features have a mean of 0 and variance of 1. This method preserves feature variance and is beneficial when data distribution doesn't conform to a specific range.

In this study, both Min-Max scaling and StandardScaler were utilized, each offering distinct advantages based on the dataset's characteristics. StandardScaler maintained the inherent distribution and variance, while Min-Max scaling ensured uniform representation across features, granting equal weightage to varying feature magnitudes.

The significance of these normalization techniques lies in their crucial roles in enhancing model performance. Addressing disparate feature scales, they contribute to a comprehensive understanding of feature relationships, empowering models to make more accurate predictions across varied datasets.

3.3.2 K-fold cross-validation (K-CV)

K-fold cross-validation (K-CV) is a widely used technique in the field of machine learning and predictive modeling for assessing the performance and generalization ability of predictive models. In K-fold cross-

validation, the original dataset is partitioned into K equal-sized subsets, or folds. The model is then trained on K-1 of the folds and validated on the remaining fold. This process is repeated K times, with each of the K folds used exactly once as the validation data. The performance of the model is then averaged over the K iterations to obtain a robust estimate of its predictive capability.

It is highlighted that in K-fold cross-validation, each estimate of a parameter is calculated based on a part of the dataset, leading to an upward bias in the cross-validation estimate of the prediction error. This emphasizes the importance of understanding the potential biases and limitations associated with K-fold cross-validation when evaluating predictive models (Fushiki, 2011).

Furthermore, it is discussed the pitfalls and challenges in selecting and assessing regression and classification models using cross-validation techniques. They emphasized the impact of high variance in models generated by current state-of-the-art methods, which can render them unsuitable for practical applications. This underscores the significance of robust cross-validation techniques such as K-fold cross-validation in addressing the variance issue and providing reliable estimates of model performance (Damjan Krstajic, 2014).

A simpler version of K-fold cross-validation exist, where the data are randomly partitioned into K equal-size subsets. This highlights the flexibility and applicability of K-fold cross-validation in various machine learning tasks, making it a versatile and widely adopted technique for model evaluation and selection (Yongli Zhang, 2015).

The K-fold cross-validation is a fundamental technique for assessing the performance of predictive models, providing a robust estimate of their generalization ability. Understanding the potential biases and limitations associated with K-fold cross-validation is crucial for ensuring the reliability of model evaluation and selection in machine learning and predictive modelling (figure 3-1).

Figure 3-1: K-Fold Cross Validation Concept with K iterations across the Test & Train datasets

3.3.3 Hyperparameter tuning

Hyperparameter tuning is a crucial process in machine learning, aiming to identify the optimal values for hyperparameters that significantly influence a model's behavior. Unlike parameters learned during training, hyperparameters are external settings that impact a model's learning, generalization, and predictive capabilities. The objective of hyperparameter tuning is to find the combination of hyperparameter values that maximizes a model's performance on a given task. This process often involves exploring a range of hyperparameter values using techniques such as grid search, random search, or Bayesian optimization.

Grid search is a widely used hyperparameter tuning technique that involves evaluating the model's performance for each combination of hyperparameter values within a predefined grid. On the other hand, random search involves randomly sampling hyperparameter values and evaluating the model's performance for each sampled combination. Bayesian optimization leverages probabilistic models to efficiently search for the optimal hyperparameter values, iteratively selecting values based on the model's performance and updating the probabilistic model (Jidesh, 2023).

Well-tuned hyperparameters can significantly enhance a model's predictive power, prevent overfitting, improve generalization, and optimize performance. This is particularly crucial for complex models and datasets with diverse characteristics, where the choice of hyperparameters can have a substantial impact on the model's behavior and performance. The impact of hyperparameter tuning on machine learning models has been studied extensively, demonstrating the significance of this process in optimizing the performance of machine learning models (Kazi Ekramul HoqueKazi, 2021).

Hyperparameter tuning is a fundamental aspect of machine learning, and the choice of hyperparameter values significantly influences a model's behavior and performance. Various techniques such as grid search, random search, and Bayesian optimization are employed to explore the hyperparameter space and identify the best configuration for a given task, ultimately leading to improved predictive power, generalization, and overall performance of machine learning models.

COA is a nature-inspired algorithm that is used for solving continuous non-linear optimization problems, based on the brood parasitism of some cuckoo species. Brood parasitism is a behavior where cuckoos lay their eggs in the nests of other host birds and let them raise their offspring.

The main features of COA are:

- It starts with an initial population of cuckoos, divided into two types: mature cuckoos and eggs.
- The cuckoos compete for survival and reproduction, based on their fitness values. Some of them die or their eggs are discovered and eliminated by the host birds.
- The survived cuckoos migrate to better habitats and lay eggs. The eggs hatch and become new cuckoos.
- The algorithm converges when there is only one cuckoo society, all with the same fitness values.

COA has been applied to various engineering and scientific problems, such as parameter optimization, scheduling, control, and image processing.

3.3.4 Data Splitting

Data splitting is a fundamental step in machine learning that involves dividing the dataset into training, testing, and sometimes validation sets. These subsets serve distinct purposes in the model development process.

Training Set:

The training set plays a crucial role in enabling the machine learning algorithm to learn the patterns and relationships within the data. It represents the majority of the dataset, typically comprising around 70-80% of the total data points. By exposing the algorithm to this portion of the data, it can effectively identify the underlying structure and relationships between the predictor variables (inputs) and the target variables (outputs).

Testing Set:

Distinct from the training set, the testing set serves as an unbiased evaluator of the trained model's performance on unseen data. This set, containing the remaining 20-30% of the data, is withheld from the training process and only used after the algorithm has been finalized. By feeding the testing set's data to the trained model, we can assess its ability to generalize to new data that it has not encountered during training. This evaluation process helps identify potential overfitting issues.

The testing set is essential for evaluating the model's generalization ability, which is its ability to perform well on new data that it has not been trained on. By comparing the model's performance on the test data to its performance on the training data, we can determine how well it generalizes to unseen data. A well-generalized model will perform similarly on both the training and testing data, indicating that it has learned the underlying patterns of the data without overfitting to the specific training data.

Data splitting is a critical technique for evaluating the performance of machine learning models and ensuring that they can effectively handle unseen data and provide reliable predictions in real-world applications.

3.4 Part one (Prediction of Permeability from predicted NMR)

This section provides a comprehensive review of the methodologies and machine learning approaches utilized in Part 01.

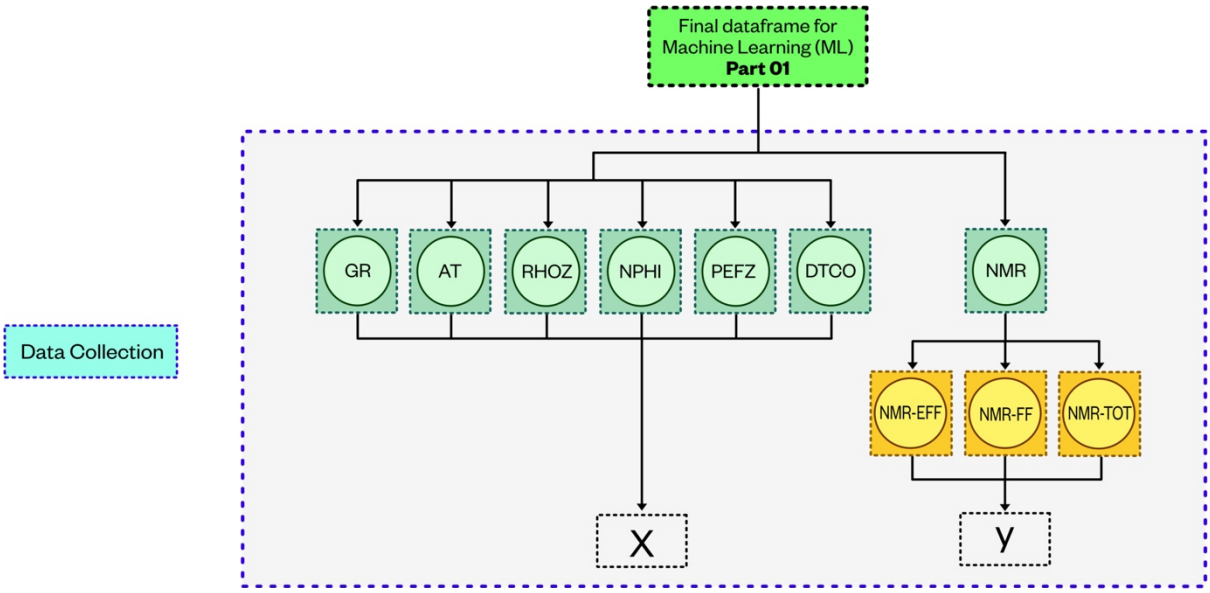

Figure 3-2: Definition of Predictors and Target for the process of Machine Learning (Part 01)

In this study, the dataset obtained from five wells has been meticulously pre-processed, segregating the dataset into two crucial components: predictors and a target variable. The predictors encompass fundamental geophysical parameters including GR (gamma ray), AT (Resistivity Log), RHOZ (bulk density), NPHI (neutron porosity), PEFZ (photoelectric factor), and DTCO (compressional wave slowness). These variables are strategically selected for their potential influence on the target variable. The target variable of interest is NMR (nuclear magnetic resonance), which holds significant importance in this context. This division sets the stage for employing predictive modeling techniques, wherein the aim is to utilize the values of these predictors to accurately forecast the NMR values. The subsequent analysis and modeling processes will be guided by the relationships and patterns discerned between these predictors and the target variable, facilitating a comprehensive understanding of the predictive capabilities within this geological context (figures 3-2, 3-3).

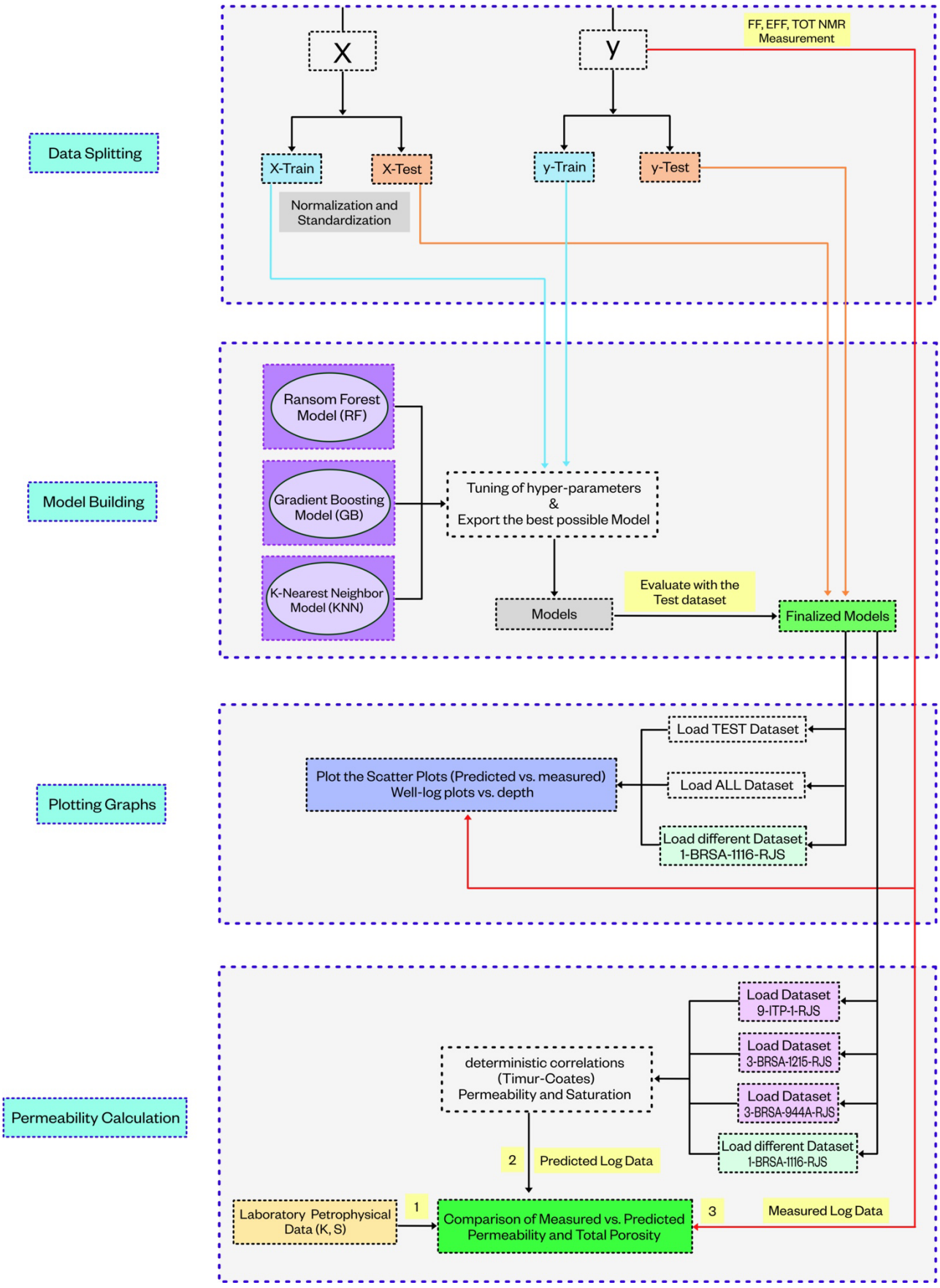

Figure 3-3: General workflow of the first part of thesis; train-test the data set contain five wells and prediction of permeability from output

3.4.1 Defining Machine Learning (ML) Model:

3.4.1.1 Random forest:

Random Forest is a powerful ensemble learning algorithm that combines the predictions of multiple decision trees to improve accuracy and reduce overfitting. It works by averaging the predictions of individual decision trees, each built using a random sample of the training data and a random subset of features. This randomness helps prevent overfitting and improves the model's stability and reliability. Random Forests are versatile and can be used for both classification and regression tasks. They have various applications in fields like natural language processing, computer vision, finance, and medicine. The algorithm has adjustable parameters such as the number of trees, tree depth, minimum sample size, bootstrapping, and feature subsampling, which can be fine-tuned to optimize model performance based on specific requirements (the architecture of Random Forest is shown in figure 3-4).

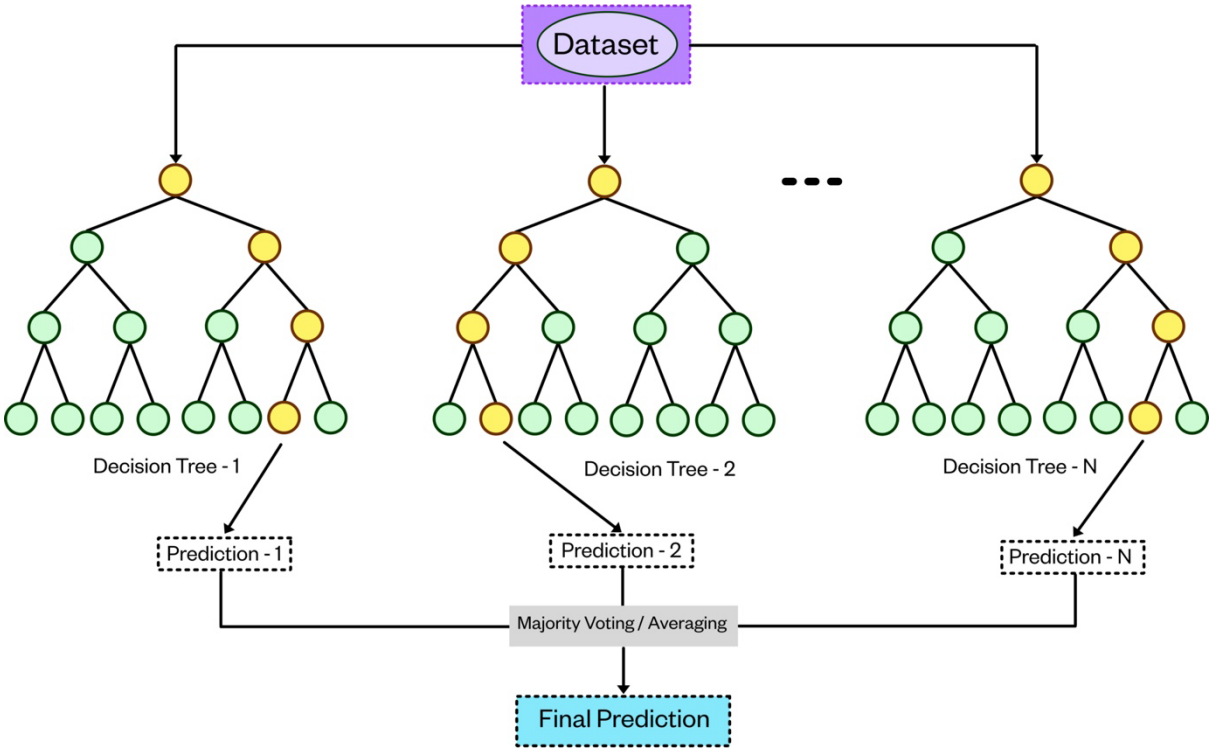

Figure 3-4: Schematic of Random Forest (RF) decision tree Architecture.

3.4.1.2 Gradient Boosting:

Gradient Boosting is a powerful ensemble method that improves predictions by sequentially correcting the errors of preceding models. Unlike Random Forests that aggregate predictions in parallel, Gradient Boosting creates an ensemble by iteratively fitting new models to the residual errors of the previous ones. This iterative learning process gradually enhances the model's predictive accuracy by minimizing a loss function. Gradient Boosting algorithms like XGBoost, LightGBM, and AdaBoost are known for

their ability to handle complex data patterns and produce highly accurate predictions. They offer adjustable parameters such as learning rate, number of trees, tree complexity, regularization parameters, and subsampling to enhance model performance. These parameters control the contribution of each model, the number of models used, the ability to capture intricate patterns, prevention of overfitting, and diversity of the model. (the architecture of Gradient Boosting is shown in figure 3-5).

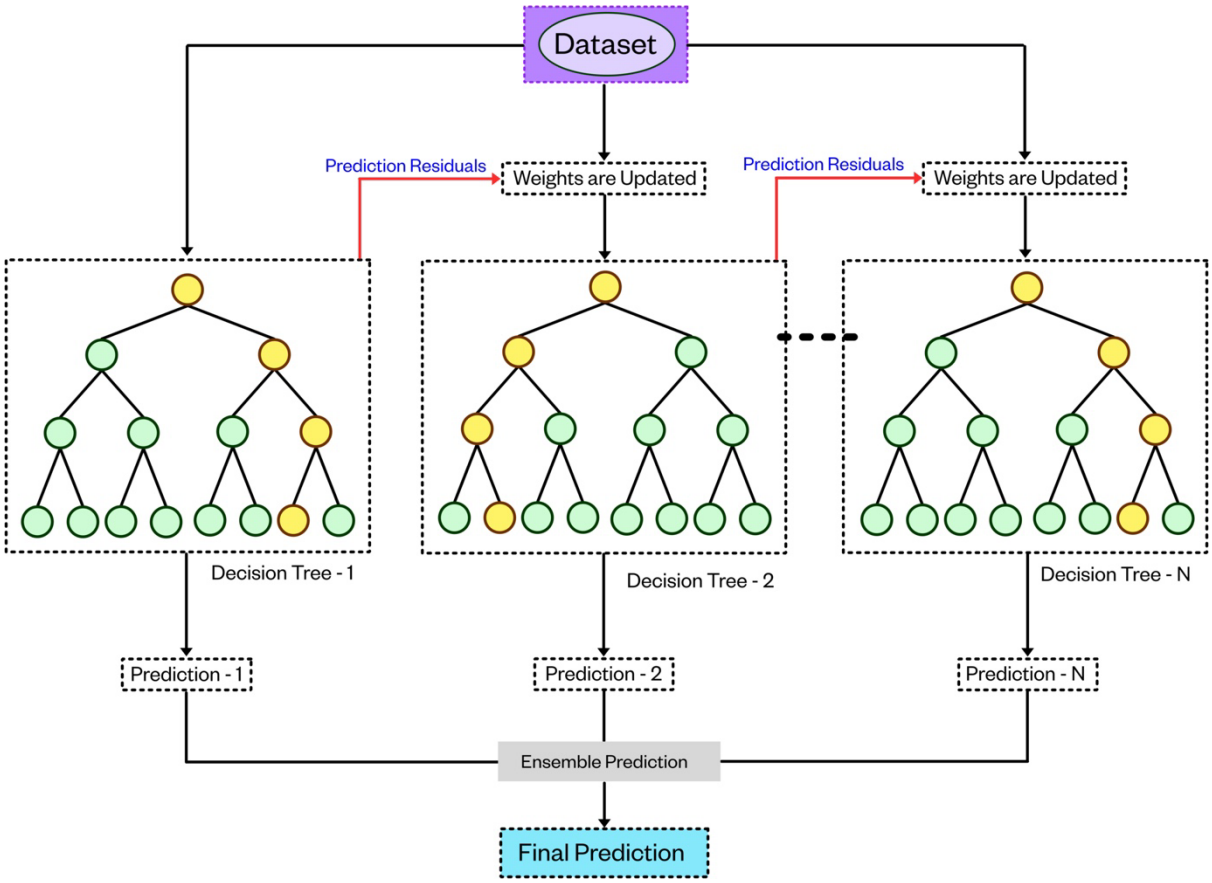

Figure 3-5: Schematic of Gradient Boosting (GB) decision tree Architecture.

3.4.1.3 k-nearest neighbor regression (k-NN):

K-nearest neighbor (k-NN) regression is a non-parametric method used in various fields, known for its simplicity and successful applications. It works by computing the k nearest neighbors of a query point and using their information to make predictions. In classification, it leverages the proximity of neighboring points to classify new data, while in regression, it estimates the response of a new data point by calculating a weighted average of the responses of its k nearest neighbors. The algorithm calculates the weighted Euclidean distance between each training point and the new data point, assigns weights to the neighbors using a kernel function, and predicts the response by computing the weighted sum of the known responses of the k nearest neighbors. The parameter k influences the flexibility of the model, with higher values resulting in smoother models (figure 3-6).

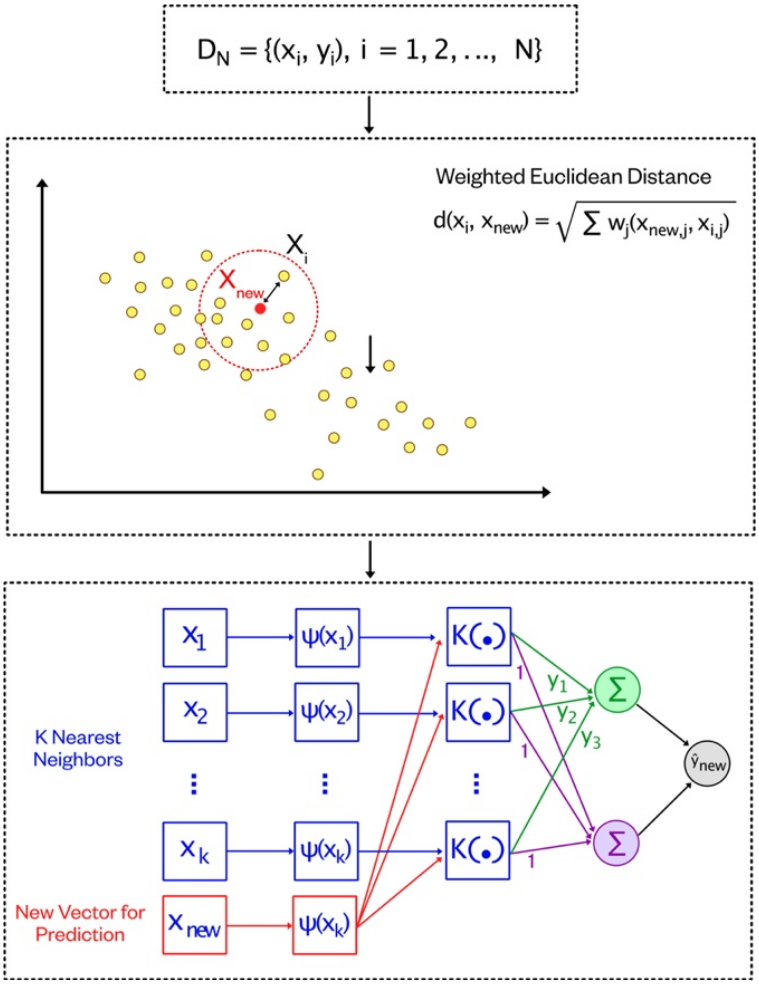

Figure 3-6: The illustration of K-Nearest Neighbor Regressor (K-NN).

The methodologies and concepts of K-NN approach is shown in Appendix 03.

3.4.2 Data Preparation and Splitting for NMR Fluid Prediction

Data preparation is a crucial step in machine learning, ensuring the quality and consistency of data to train accurate models. In this case, the data consists of geophysical log measurements (predictors) and NMR fluid parameters (target variables) collected from various boreholes.

Predictors (Input Variables), the provided data comprises six predictor variables extracted from geophysical logs:

- GR: Gamma Ray: Measures the natural radioactivity and provides lithlogy (Shale vs. sand).
- AT90: Measures resisitivity of the formation and provides water saturation.
- RHOZ: Measures bulk Density of the formation that can also provide porosity.
- NPHI: Neutron Porosity: Measures the hydrogen content of the formation and provides porosity.
- DTCO: Delta-T (Interval Transit Time P-Waves): Reflects the meachanical properties.
- PEFZ: Photoelectric Factor: Assists in determining the lithology of the formation.

These predictor variables provide valuable insights into the characteristics of the subsurface geology, which can be used to predict NMR fluid parameters.

Target Variables (Outputs), the target variables represent the NMR fluid parameters to be predicted:

- NMREFF: Effective Nuclear Magnetic Resonance: Indicates the total amount of hydrocarbon saturation in the formation.
- NMRFF: Free Fluid Volume using NMR: Measures the amount of free fluid (oil and gas) in the pores.
- NMRTOT: Total Nuclear Magnetic Resonance: Represents the total NMR signal strength, encompassing all hydrocarbon and liquid saturations.

These target variables are crucial for understanding the hydrocarbon distribution and potential reservoir quality.

Data Splitting:

To effectively train and evaluate the machine learning model, the data is split into training and testing sets using an 80/20 split.

- Training Data:

The training data consists of 8340 samples, each containing six predictor variables and corresponding values for the three target variables. This portion of the data is used to train the machine learning algorithm, allowing it to learn the patterns and relationships between the predictors and the target variables.

- Testing Data:

The testing data comprises 2086 samples, again with six predictor variables and matching values for the three target variables. This set is withheld from the training process and used to evaluate the performance of the trained model. The model's predictions on the testing data provide an unbiased assessment of its generalizability to unseen data.

- Shapes of Data:

The original data, before splitting, comprises 10,426 samples for the predictors (X) and 10,426 samples for the target variables (y). This indicates that each sample has data for all six predictors and all three target variables.

3.4.3 Hyperparameter Optimization Across Models

Random Forest:

- Ensemble Size (n_estimators): The number of decision trees in the random forest ensemble was systematically varied from 100 to 400 in increments of 50, examining how the ensemble size influenced the overall performance of the model.
- Decision Tree Complexity (max_depth): The maximum depth of each decision tree in the ensemble was explored across a range of values from 5 to 25, aiming to balance the model's ability to capture complex patterns without overfitting to the training data.

- Loss Function (criterion): The default criterion for evaluating splits in the decision trees, 'squared_error', was investigated to determine its effectiveness in minimizing the loss function and improving the model's predictive accuracy.

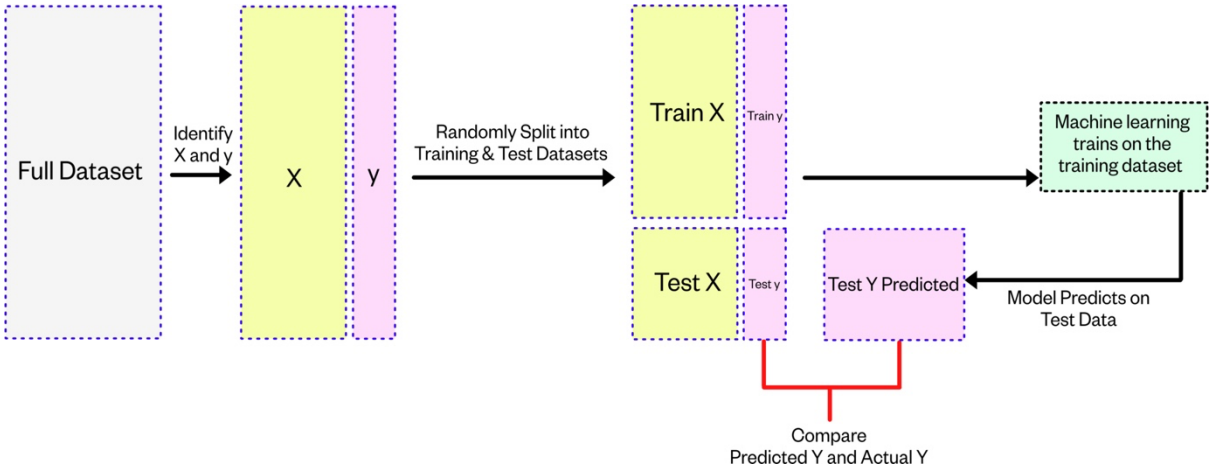

Figure 3-7: Train-test splitting configuration

Gradient Boosting:

- Ensemble Size, Tree Depth, and Learning Rate: The number of trees ('n_estimators'), maximum depth of each tree ('max_depth'), and learning rate ('learning_rate') were jointly tuned to optimize the performance of the gradient boosting algorithm. The 'learning_rate' parameter controls the step size in the gradient descent optimization process, ranging from 0.1 to 0.6 to find the optimal balance between convergence speed and overfitting tendency.
- Loss Function (criterion): Similar to Random Forest, the default criterion for evaluating splits in the gradient boosting trees, 'squared_error', was evaluated to assess its impact on the model's predictive performance.

K-Nearest Neighbors (KNN):

When working with the K-nearest neighbors (KNN) algorithm, several hyperparameters can be adjusted to optimize its performance. The number of nearest neighbors (n_neighbors) determines the size of the neighborhood considered for classification or regression tasks. Neighbor weighting (weights) assigns weights to each neighbor in the KNN calculation, influencing the model's behavior. The distance metric (p) measures the similarity between points and affects the model's performance. The leaf size (leaf_size) determines the granularity of the representation and impacts the model's generalization ability. Different algorithms (algorithm) can be used for optimizing the nearest neighbor search. By exploring and selecting appropriate values for these hyperparameters, the KNN model can be fine-tuned to achieve better results for specific datasets and tasks.

The exploration of hyperparameters in machine learning algorithms emphasizes the significance of optimizing model configurations for optimal performance and generalization. By systematically varying the hyperparameters, the study aimed to identify configurations that enhance predictive accuracy and

robustness across different parameter spaces, leading to effective models with consistent performance in various applications.

3.4.4 Well-log data organisation (Test and all datasets)

The effective utilization of machine learning models hinges on a meticulous selection of training, validation, and test wells. These wells represent distinct data subsets, each fulfilling a crucial role in model development and evaluation.

The training well serves as the cornerstone of model building, providing the data the model utilizes to decipher the intricate relationships between input and target variables. It acts as a microcosm where the model observes patterns, identifies correlations, and develops predictive mechanisms.

The validation well, distinct from the training well, plays a pivotal role in model selection and hyperparameter tuning. Representing unseen data, it provides an unbiased assessment of the model's performance on novel data. By comparing the model's predictions on the validation well to the actual values, we can evaluate its generalizability and identify optimal hyperparameter settings.

The test well, entirely separate from the model-building process, assumes the role of an independent measure of the model's generalizability. Representing entirely new data, never encountered during training, it serves as the ultimate arbiter of the model's true predictive power. By comparing the model's predictions on the test well to the actual values, we can assess its ability to generalize beyond the training data and apply its learnings to real-world scenarios.

A summarized view of the well-log predictors and outputs of part 01 are shown in figure 3-8:

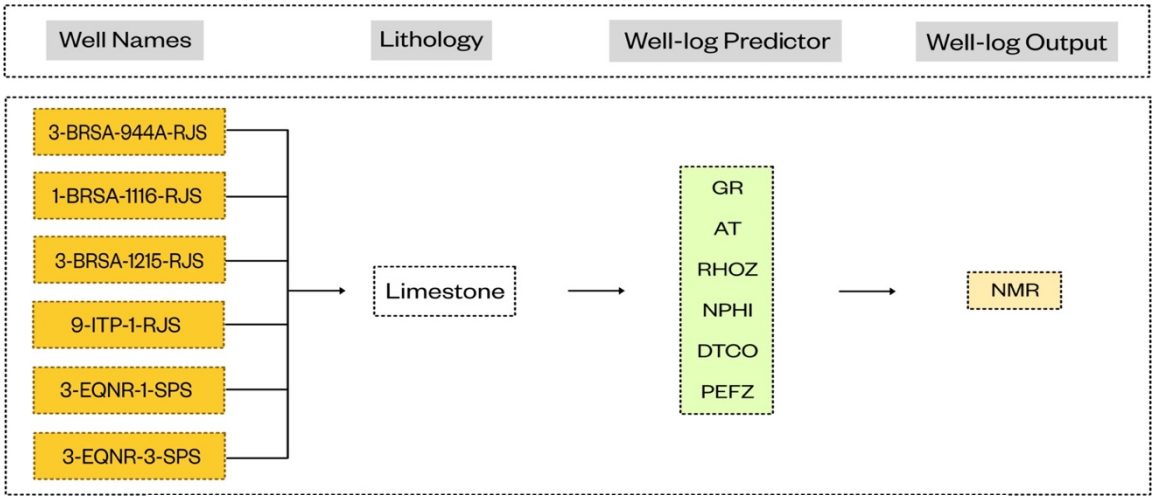

Figure 3-8: General description of wells, geology types, predictors, and target data (part

3.4.5 The Timur-Coates model; Permeability & Saturation estimation

The Timur-Coates model is a widely used method for estimating permeability in petrophysics. It is important to note that the Timur-Coates model is one of the major methods for estimating permeability, alongside the Schlumberger-Doll-Research (SDR) model (Jun Yoneda, 2022). This model has been applied in various geological settings, including the estimation of permeability in tight reservoirs (Razieh Solatpour, 2018), mudrocks, and fine-grained sandstones in coal reservoirs (Na Zhang, 2018). Additionally, it has been used for estimating the permeability coefficient of artificial clayey soil (Zhen Lu, 2020) and in the construction of a petrophysical model for a heterogeneous reservoir in the South China Sea (Xing Lei, 2021).

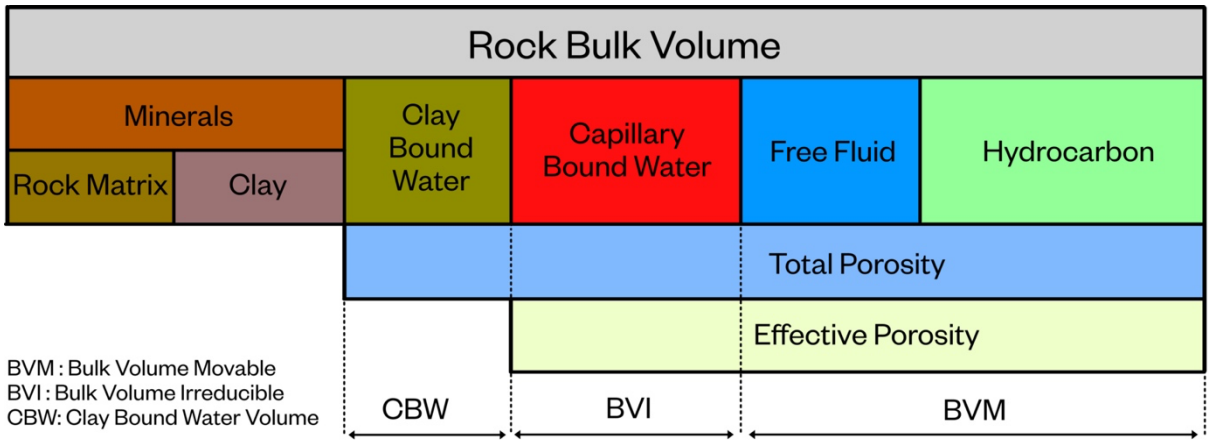

Figure 3-9: Schematic representation of the constituents of a rock; Definition of different porosities; definition of BVM, BVI, CBW.

The Timur-Coates model is particularly valuable due to its application in estimating permeability from nuclear magnetic resonance (NMR) data (figure 3-9). It has been used to calculate reservoir porosity and permeability from NMR logs, providing a non-destructive and cost-effective alternative to traditional core analysis and well testing methods. The Timur-Coates model plays a crucial role in the field of petrophysics, offering a valuable tool for estimating permeability in diverse geological settings. Its utilization in NMR analysis and its cost-effectiveness makes it an essential method for characterizing reservoir properties and evaluating the potential for hydrocarbon production. The equations 3-1 to 3-8 show the Timur-Coates correlations for petrophysical properties estimation:

$$\phi_t = Movable\ Volume + Immovable\ Volume = BVM + (BVI + CBW)$$

Equation 3-1

$$\phi_t = BVM + (BVI + CBW)$$

Equation 3-2

$$CBW = \phi_t - \phi_e$$

Equation 3-3

$$BFE = \phi_e - F$$

Equation 3-4

$$BF = \phi_t - FF$$

Equation 3-5

$$Swi_t = \left(1 - \frac{\phi_t - FF}{\phi_t}\right) \times 100 = \left(\frac{BF}{\phi_t}\right) \times 10$$

Equation 3-6

$$Swi_e = \left(1 - \frac{FF}{\phi_e}\right) \times 100 = \left(\frac{\phi_e - FF}{\phi_e}\right) \times 100 = \left(\frac{BFE}{\phi_e}\right) \times 100$$

Equation 3-7

$$K = a \times \left(\frac{FF}{BF}\right)^b \times {\phi_t}^c \quad , \qquad a = 10000, \qquad b = 2, \qquad c = 4$$

Equation 3-8

3.5 Part two (Prediction of Permeability directly from well-log data)

This section provides a comprehensive review of the methodologies and machine learning approaches utilized in Part 02 (figure 3-10).

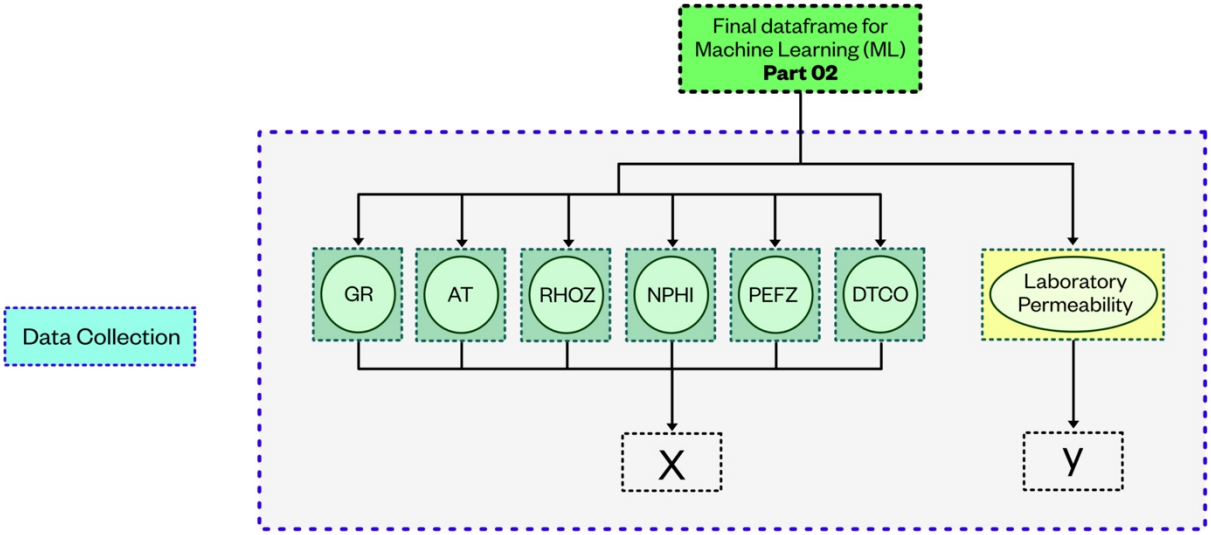

Figure 3-10:Definition of Predictors and Target for the process of Machine Learning (Part 02)

In this study, the dataset obtained from four wells has been meticulously pre-processed, segregating the dataset into two crucial components: predictors and a target variable. The predictors encompass fundamental geophysical parameters including GR (gamma ray), AT (acoustic travel time), RHOZ (bulk density), NPHI (neutron porosity), PEFZ (photoelectric factor), and DTCO (compressional wave slowness). These variables are strategically selected for their potential influence on the target variable. The target variable of interest is permeability (K) from laboratory data, which holds significant importance in this context. This division sets the stage for employing predictive modelling techniques, where the aim is to utilize the values of these predictors to accurately forecast the Permeability values. The subsequent analysis and modelling processes will be guided by the relationships and patterns discerned between these predictors and the target variable, facilitating a comprehensive understanding of the predictive capabilities within this geological context (figure 3-11).

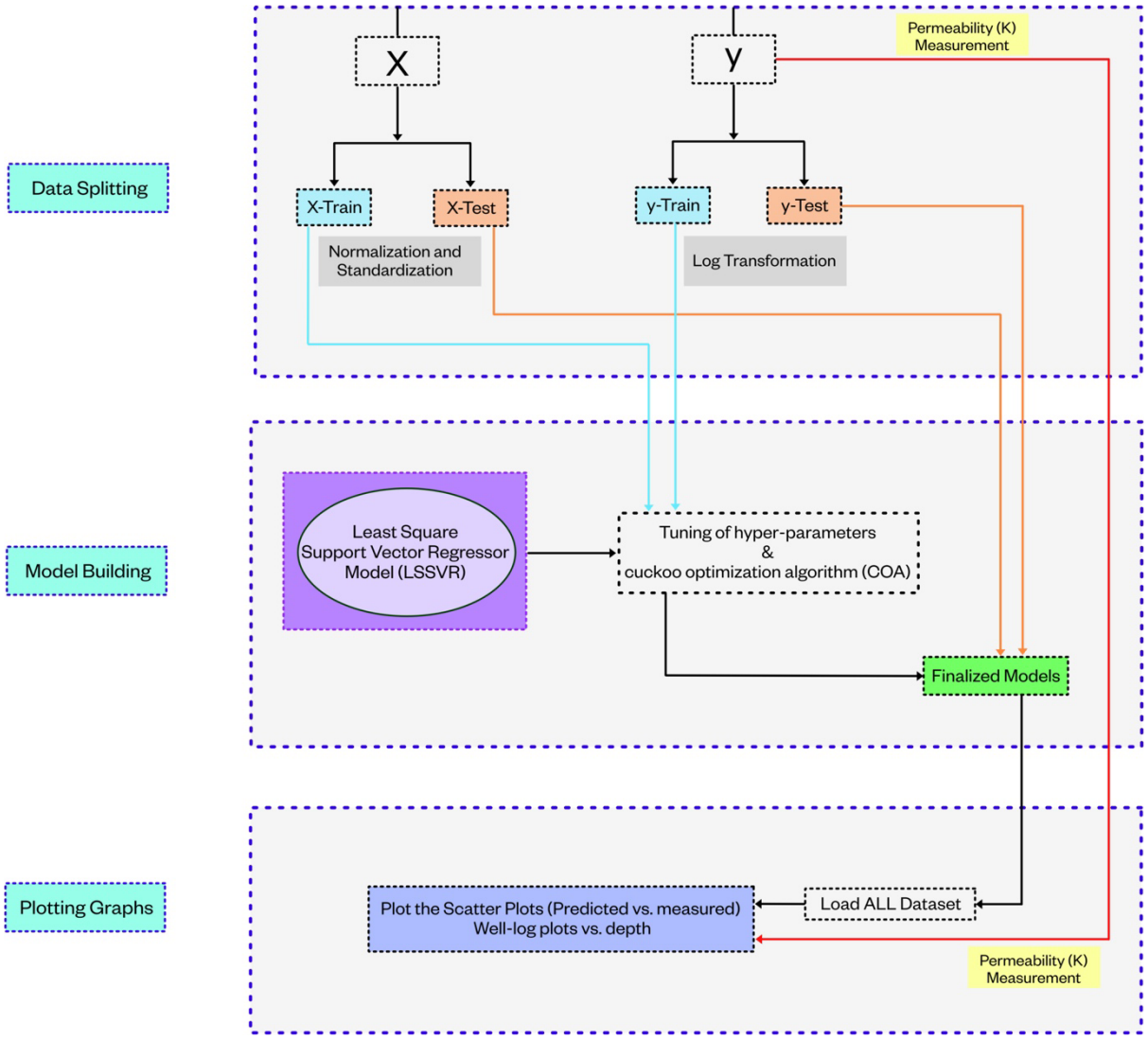

Figure 3-11: General workflow of the second part of thesis; train-test the data set contain five wells and prediction of permeability from output

3.5.1 Defining Machine Learning (ML) Model:

Selecting suitable machine learning (ML) algorithms for the analysis of a cleaned well-log dataset is a crucial step in the data analysis process. In this part of the thesis, we implement LSSVM approach.

Least Square Support vector machine (LSSVM)

Least Square Support Vector Machine (LSSVM) has gained attention in various fields due to its superior prediction accuracy and efficiency. The LSSVM has been successfully utilized in the petroleum industry to predict permeability impairment caused by scale deposition during a water injection process. It has been applied in control theory for induction motor drives in electric vehicles and integrated with other models for time series prediction. LSSVM is a reformulation of support vector machines (SVM) and its

versatility and potential for addressing complex prediction and estimation challenges make it a widely adopted method (the architecture of LSSVM methodologies is show in figure 3-12).

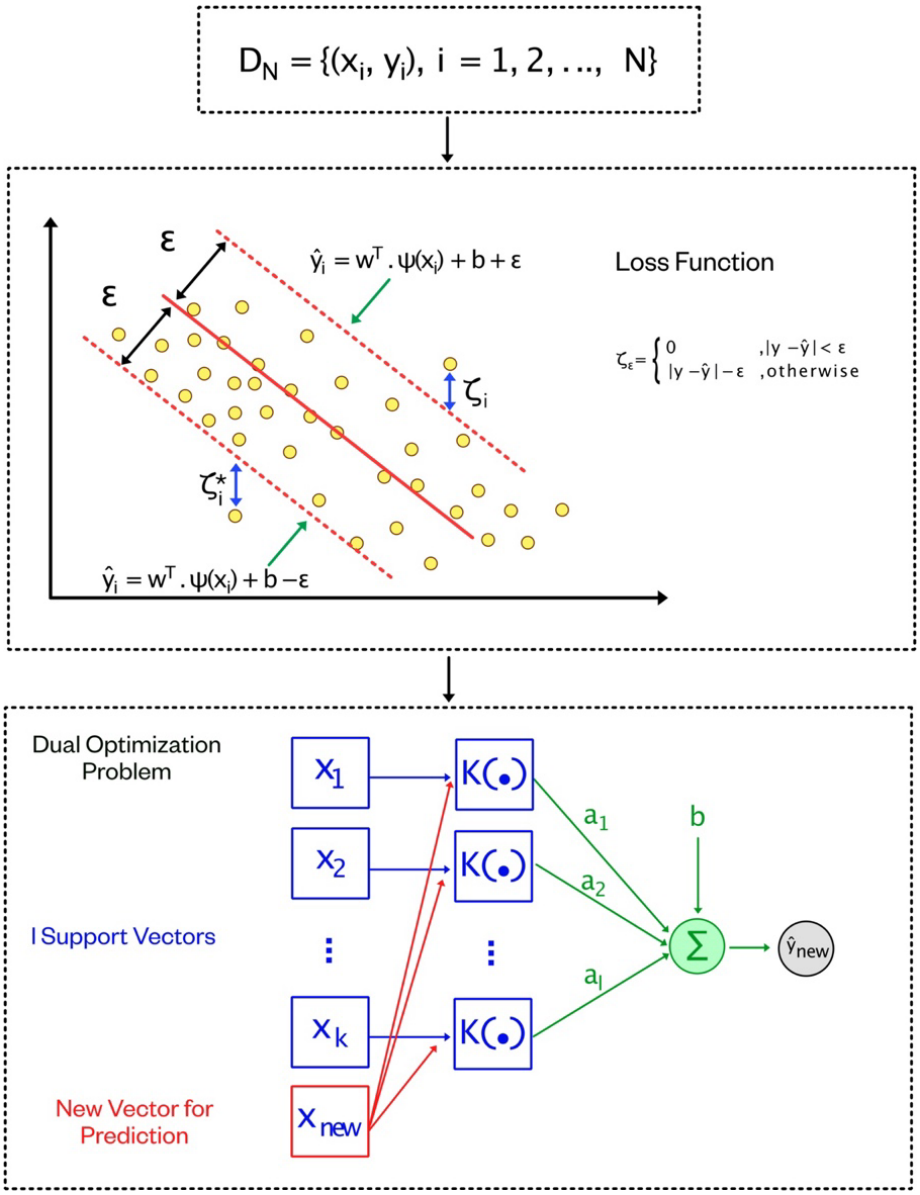

Figure 3-12: The illustration of Support Vector Machine (SVM).

The methodologies and concepts of SVM approach is shown in Appendix 03.

3.5.2 Data Preparation and Splitting for Permeability Prediction

Data preparation is a crucial step in machine learning, ensuring the quality and consistency of data to train accurate models. In this case, the data consists of geophysical log measurements (predictors) and NMR fluid parameters (target variables) [peameability] collected from various boreholes.

Predictors (Input Variables), the provided data comprises six predictor variables extracted from geophysical logs:

- GR: Gamma Ray: Measures the natural radioactivity and provides lithlogy (Shale vs. sand).
- AT90: Measures resisitivity of the formation and provides water saturation.
- RHOZ: Measures bulk Density of the formation that can also provide porosity.
- NPHI: Neutron Porosity: Measures the hydrogen content of the formation and provides porosity.
- DTCO: Delta-T (Interval Transit Time P-Waves): Reflects the meachanical properties.
- PEFZ: Photoelectric Factor: Assists in determining the lithology of the formation.

These predictor variables provide valuable insights into the characteristics of the subsurface geology, which can be used to predict NMR fluid parameters.

Target Variables (Outputs), the target variables represent the NMR fluid parameters to be predicted:

- Permeability (K): Permeability from laboratory data Indicates the ability to flow the hydrocarbon

These target variables are crucial for understanding the hydrocarbon distribution and potential reservoir quality.

Data Splitting:

For the machine learning model, the data was split into training and testing sets using a 70/30 split.

- Training Data:

The training data comprises 191 samples, each containing six predictor variables: RHOZ, GR, AT90, NPHI, DTCO, and PEFZ. These variables are used to train the machine learning algorithm, enabling it to learn the relationships between the predictors and the target variable 'k'.

- Testing Data:

The testing data consists of 82 samples, also with the same six predictor variables. This data is held back from the training process and employed to evaluate how well the trained model performs. The model's predictions on the testing data offer an impartial assessment of its ability to generalize to new, unseen data.

- Shapes of Data:

The original dataset contained 273 samples for the predictors (X) and 273 samples for the target variable 'k'. Each sample in this dataset included data for all six predictors and the single target variable.

3.5.3 Hyperparameter Optimization Across Models

Least Square Support Vector Machine (LSSVM):

- Gamma Variance (gamma): The gamma parameter, set at 1.0, plays a pivotal role in the LSSVR model, influencing the width of the radial basis function kernel. This parameter governs the influence of individual training samples in the model. The exploration of this parameter aims to ascertain its impact on the model's performance across various datasets.

- Regularization Parameter (C): The regularization parameter C, set at 100.0, is another crucial factor in the LSSVR model. It controls the trade-off between achieving a low training error and minimizing model complexity. Systematic variations in the C parameter were executed to evaluate its influence on model performance.
- Kernel Function (kernel): The radial basis function (RBF) kernel, chosen for this LSSVR implementation, offers flexibility in capturing non-linear relationships between the predictors and the target variable. The 'rbf' kernel type was utilized to build the LSSVR model and its effectiveness was assessed in modeling complex patterns within the data.

3.5.4 Well-log data organisation (Test and all datasets)

The training subset serves as the foundation for constructing the LSSVR model, enabling it to discern intricate patterns between input predictors and the target variable. Within this subset, the model deciphers correlations, learns predictive mechanisms, and establishes its understanding of the data.

Distinct from the training set, the validation subset assumes a crucial role in selecting the optimal LSSVR model configuration. It mirrors unseen data, allowing an unbiased evaluation of the model's performance. By comparing the model's predictions on this subset with actual values, we gauge its generalizability and identify the most effective parameter settings.

As an independent entity from the model development phase, the test subset acts as a litmus test for the model's real-world predictive prowess. Comprising entirely new data, it remains unseen during model training or validation. Assessing the model's predictions on this subset against actual values provides a final measure of its ability to generalize beyond the training data, ensuring its applicability in real-world scenarios.

A summarized view of the well-log predictors and outputs of part 02 is shown in figure 3-13:

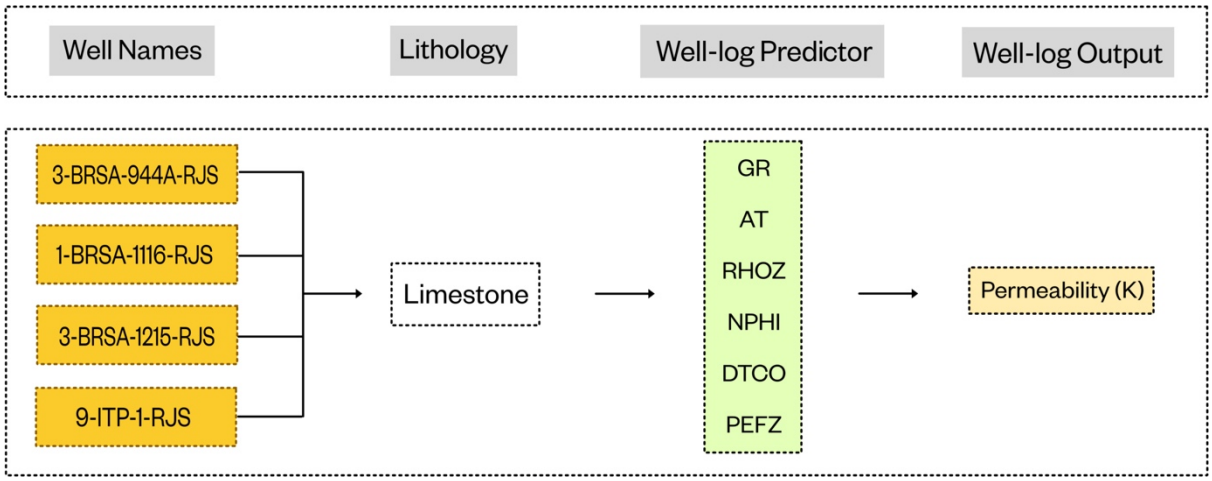

Figure 3-13: General description of wells, geology types, predictors, and target data (part 02)

4 Methodology Application & Outcomes

4.1 Part one

4.1.1 Evaluation of different model performances using the test dataset

4.1.1.1 Scatter plots

These scatter plots represent different porosities (free fluid, effective, and total) of predicted and measured which is show in figures 4-1, 4-2, 4-3:

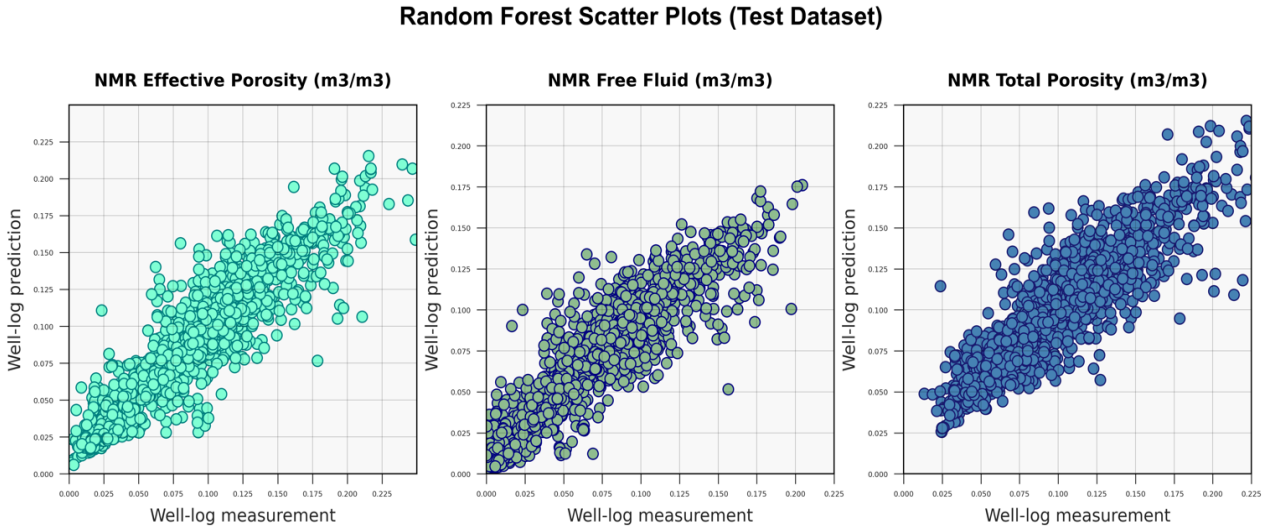

Figure 4-1: Scatter plots of predicted versus measured NMR porosity, for the RF Model, applied to the test dataset of the original dataset contain whole five wells; Left plot: NMR Effective Porosity (m3/m3). Middle plot: NMR Free Fluid (m3/m3). Right plot: NMR Total Porosity (m3/m3).

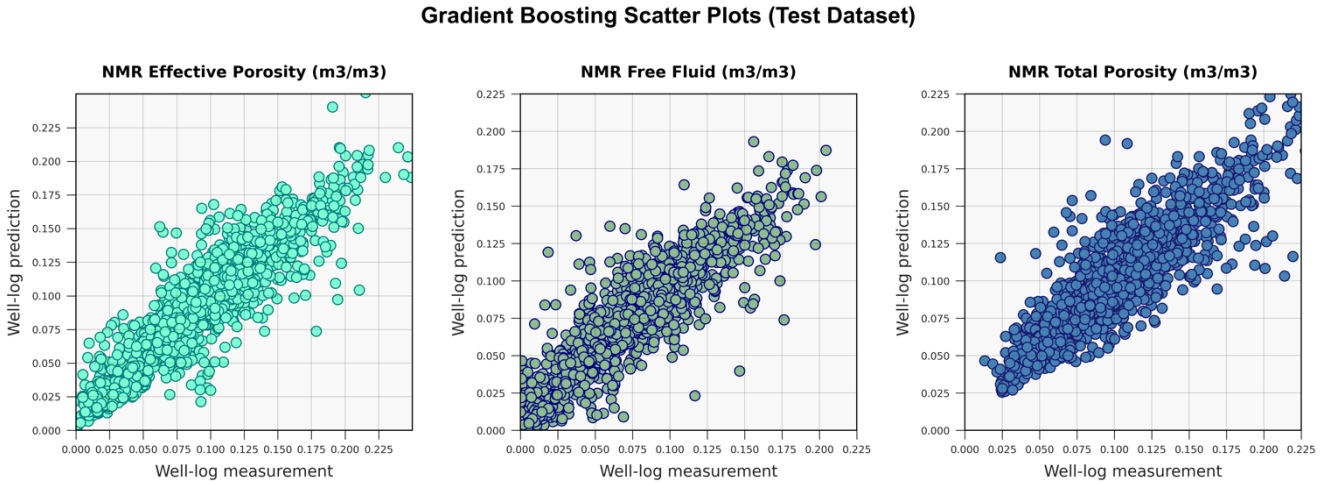

Figure 4-2: Scatter plots of predicted versus measured NMR porosity, for the GB Model, applied to the test dataset of the original dataset contain whole five wells: Left plot: *NMR Effective Porosity (m3/m3). Middle plot: NMR Free Fluid (m3/m3). Right plot: NMR Total Porosity (m3/m3).*

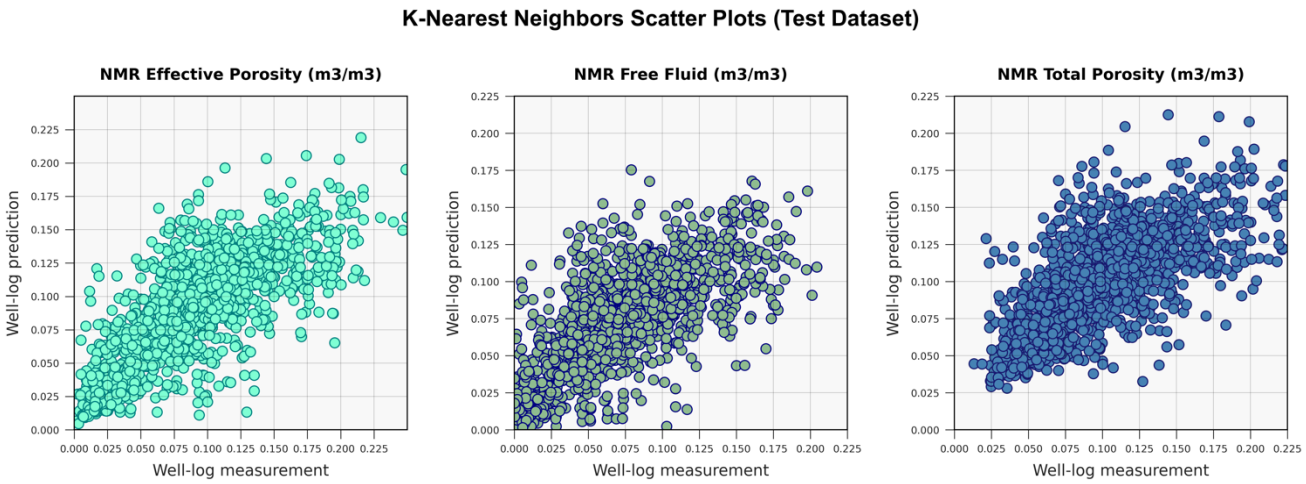

Figure 4-3: Scatter plots of predicted versus measured NMR porosity, for the K-NN Model, applied to the test dataset of the original dataset contain whole five wells. Left plot: NMR Effective Porosity (m3/m3). Middle plot: NMR Free Fluid (m3/m3). Right plot: NMR Total Porosity (m3/m3).

- **Model Deviation:**

Scatter plots are graphical representations showcasing the variance or deviation between the predicted values from a model and the actual measured data points. When these data points form a straight line at a 45-degree angle, aligning with the ideal diagonal, it signifies a strong accuracy between the model predictions and the observed data.

- **Test Dataset Density:**

The model is developed based on an 80/20 splitting ratio, dividing the dataset into training (80%) and testing (20%) subsets. Consequently, the resulting test dataset might be relatively smaller, leading to a less densely populated scatter plot. This reduced density can affect the clarity and strength of the observed trends, potentially resulting in a weaker trend line representation within the scatter plot.

- **Distribution of Porosity Types and Their Positioning:**

Within the dataset, total porosity tends to possess higher numerical values compared to effective and free fluid porosities. Consequently, when visualized on the scatter plot, the data points representing total porosity often cluster towards the higher value range. This tendency places total porosity data points on the right side of the plot due to their higher magnitude, delineating their dominance in the dataset distribution.

Visually, the Random Forest (RF) and Gradient Boosting (GB) show a better performance in comparison to K-nearest Neighbours (K-NN)

4.1.1.2 Well-log Plots

These plot showcases the comparison between predicted and measured porosities encompassing effective, free fluid, and total porosities. As models inherently contain some margin of error and imprecision, the predicted porosity doesn't perfectly align with the measured well-log data. You can find these plots in figures 4-4, 4-5, and 4-6.

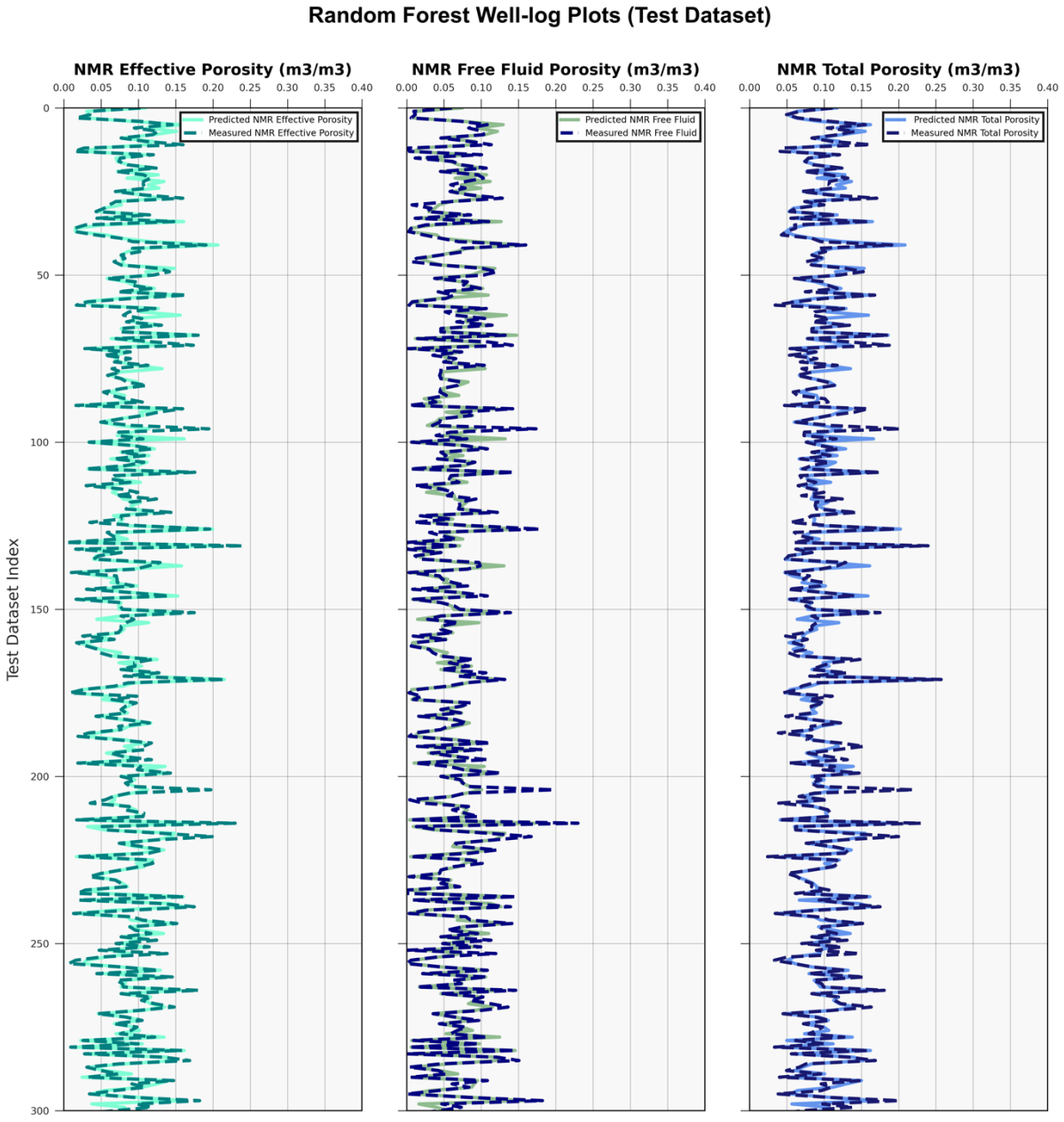

Figure 4-4: Comparing the match between the predicted and measured NMR porosity, for the RF Model, across the Test Dataset Index of the original dataset contain whole five wells. Left track: Predicted and Measured NMR Effective Porosity (m3/m3). Middle track: Predicted and Measured NMR Free Fluid (m3/m3). Right track: Predicted and Measured NMR Total Porosity (m3/m3).

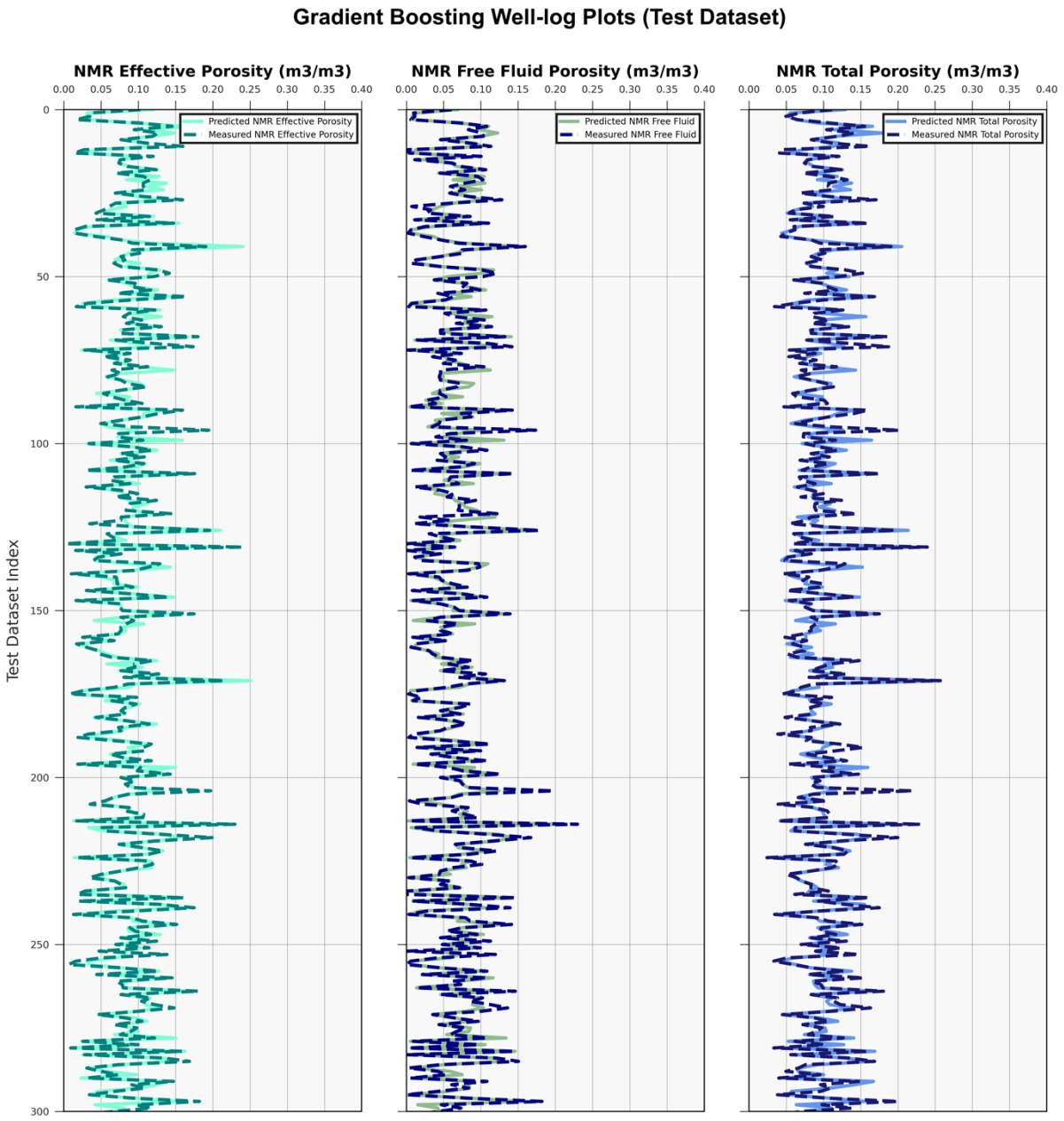

Figure 4-5: Comparing the match between the predicted and measured NMR porosity, for the GB Model, across the Test Dataset Index of the original dataset contain whole five wells. Left track: Predicted and Measured NMR Effective Porosity (m3/m3). Middle track: Predicted and Measured NMR Free Fluid (m3/m3). Right track: Predicted and Measured NMR Total Porosity (m3/m3).

Figure 4-6: Comparing the match between the predicted and measured NMR porosity, for the K-NN Model, across the Test Dataset Index of the original dataset contain whole five wells. Left track: Predicted and Measured NMR Effective Porosity (m3/m3). Middle track: Predicted and Measured NMR Free Fluid (m3/m3). Right track: Predicted and Measured NMR Total Porosity (m3/m3).

There are a number of factors that can contribute to the inaccuracy of porosity models. These include:

- The inherent complexity of the geology: The rock formations that make up a reservoir can be very complex and heterogeneous, which makes it difficult to create a model that accurately captures all of the variations in porosity.
- The limitations of the data: The data that is used to train the model may not be perfect, and it may not be representative of all the conditions that exist in the reservoir.

4.1.2 Evaluation of different model performances using the entire dataset

4.1.2.1 Scatter plots

By applying entire dataset on different models for prediction of porosities, more correlated and clearer trend between predicted and measured well-log data have been presented. Due to the methodology of gradient boosting, a more correlated trend is shown. The gradient boosting model, which combines weak learners into a strong learner by optimizing a loss function, performs better than the other models in terms of correlation coefficient and mean squared error. This indicates that the gradient boosting model can capture the complex relationship between the input features and the output variable more effectively. You can find these plots in figures 4-7, 4-8, and 4-9.

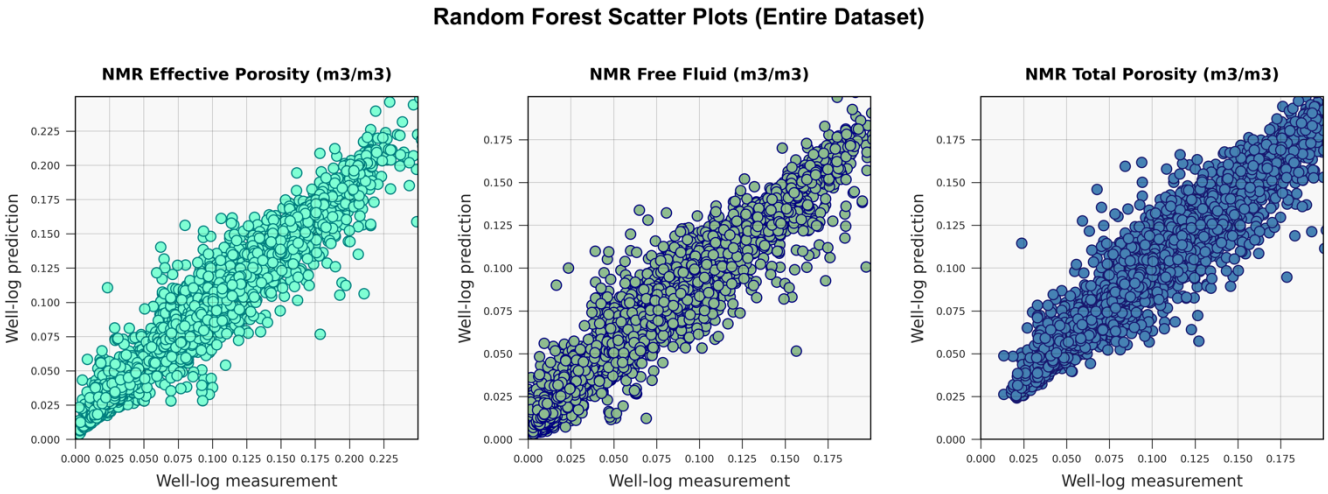

Figure 4-7: Scatter plots of predicted versus measured NMR porosity, for the RF Model, applied to the entire dataset of the original dataset contain whole five wells: Left plot: NMR Effective Porosity (m3/m3). Middle plot: NMR Free Fluid (m3/m3). Right plot: NMR Total Porosity (m3/m3).

In the scatter plots of K-nearest neighbour model, there are some over-fittings and this will generally lead to lower accurate model. Overfitting occurs when a model learns the training data too well, capturing not just the underlying patterns but also the noise or random fluctuations. In the context of scatter plots in K-nearest neighbor (KNN) models, overfitting might be observed when the decision boundary (or separation between classes) becomes overly complex and fits intricacies in the training data that don't generalize well to new, unseen data.

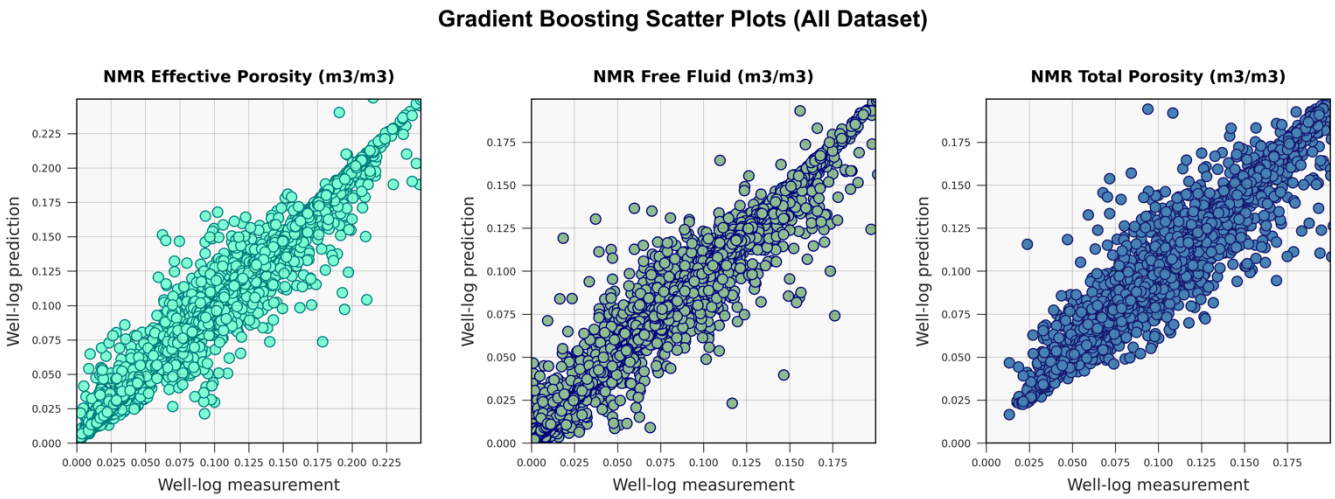

Figure 4-8: Scatter plots of predicted versus measured NMR porosity, for the GB Model, applied to the entire dataset of the original dataset contain whole five wells: Left plot: NMR Effective Porosity (m3/m3). Middle plot: NMR Free Fluid (m3/m3). Right plot: NMR Total Porosity (m3/m3).

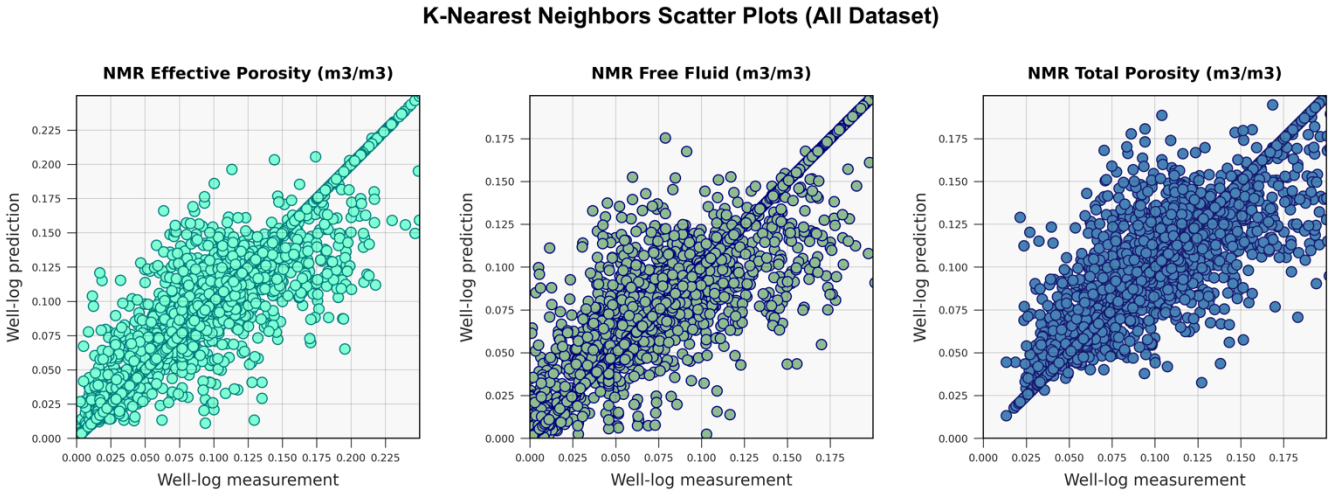

Figure 4-9: Scatter plots of predicted versus measured NMR porosity, for the K-NN Model, applied to the entire dataset of the original dataset contain whole five wells: Left plot: NMR Effective Porosity (m3/m3). Middle plot: NMR Free Fluid (m3/m3). Right plot: NMR Total Porosity (m3/m3).

lower accuracy. It's essential to balance model complexity and generalization by tuning hyperparameters, like the number of neighbors (K) in KNN, or using techniques like cross-validation and regularization to prevent overfitting and build a more accurate model.

4.1.2.2 Well-log Plots

Using the entire dataset for modeling presents a comprehensive view of the predictive capabilities across the entire range of available data. While this approach can provide insights into overall trends and patterns, it carries risks. Over-reliance on the entire dataset might lead to overfitting, where the model becomes too tailored to the specific intricacies of the data, potentially hindering its ability to generalize well to new, unseen cases. The plots of measured and predicted porosities can be found in figures 4-10, 4-11, and 4-12.

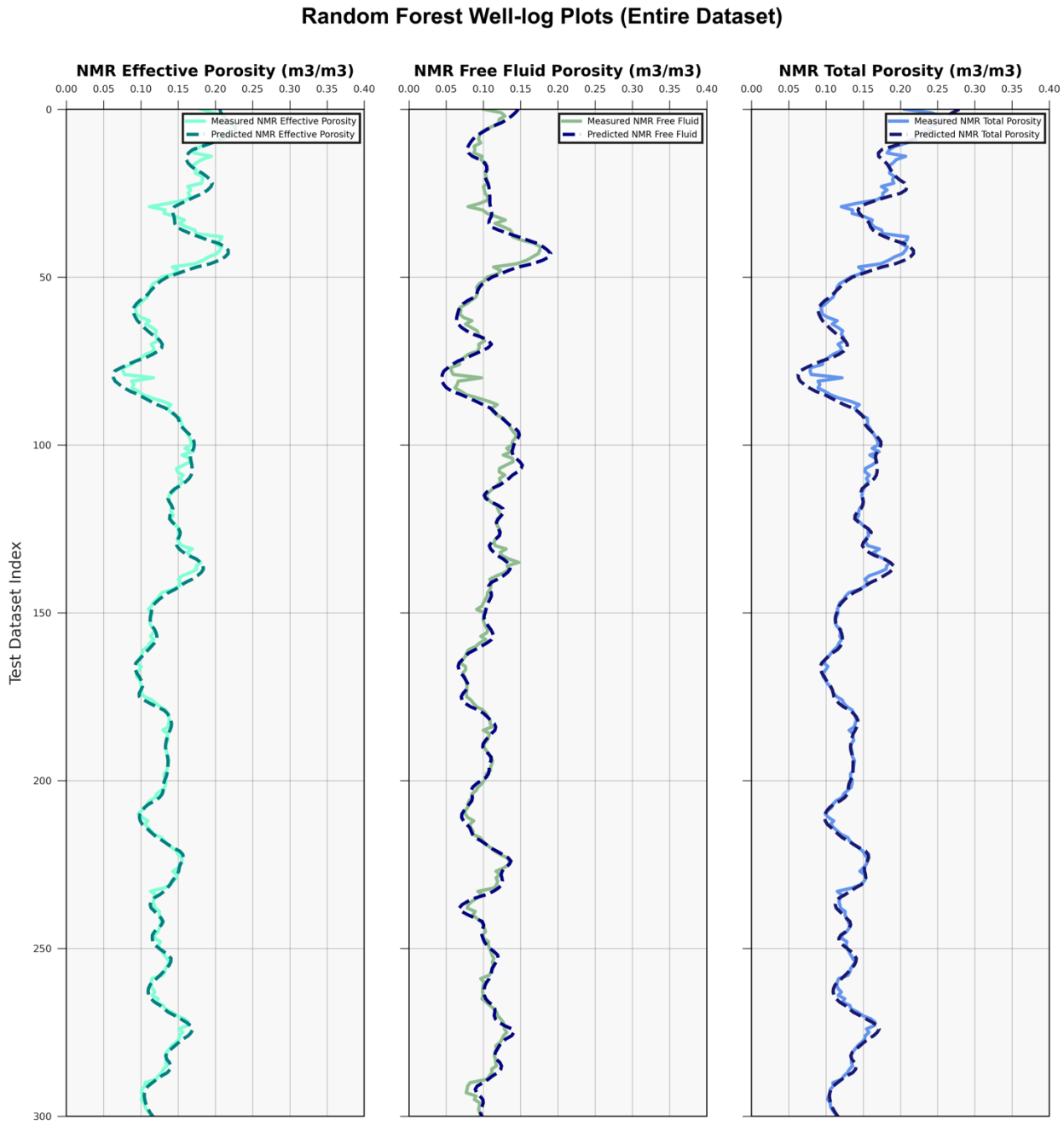

Figure 4-10: Comparing the match between the predicted and measured NMR porosity, for the RF Model, across the Entire dataset index of the original dataset contain whole five wells. Left track: Predicted and Measured NMR Effective Porosity (m3/m3). Middle track: Predicted and Measured NMR Free Fluid (m3/m3). Right track: Predicted and Measured NMR Total Porosity (m3/m3).

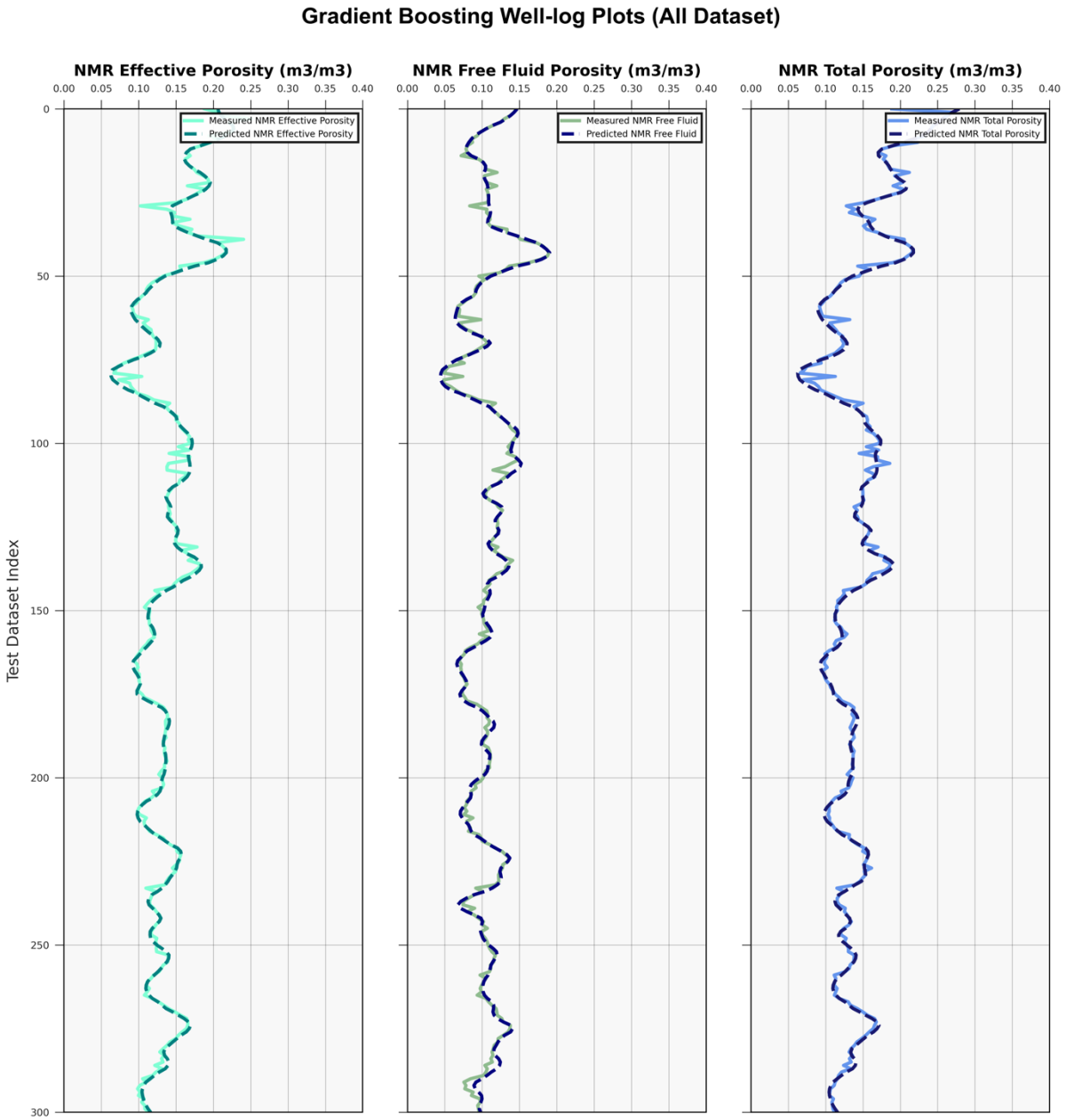

Figure 4-11: Comparing the match between the predicted and measured NMR porosity, for the GB Model, across the Entire dataset index of the original dataset contain whole five wells. Left track: Predicted and Measured NMR Effective Porosity (m3/m3). Middle track: Predicted and Measured NMR Free Fluid (m3/m3). Right track: Predicted and Measured NMR Total Porosity (m3/m3).

The discrepancy observed between predicted and measured porosities underscores the inherent limitations of models in capturing the complexities of geological formations. Highlighting these limitations is crucial; it acknowledges the margin of error and imprecision that inherently exists in predictive models. By acknowledging these limitations, researchers can refine models and data collection strategies, seeking more robust approaches to improve predictive accuracy and account for variability in porosity estimation.

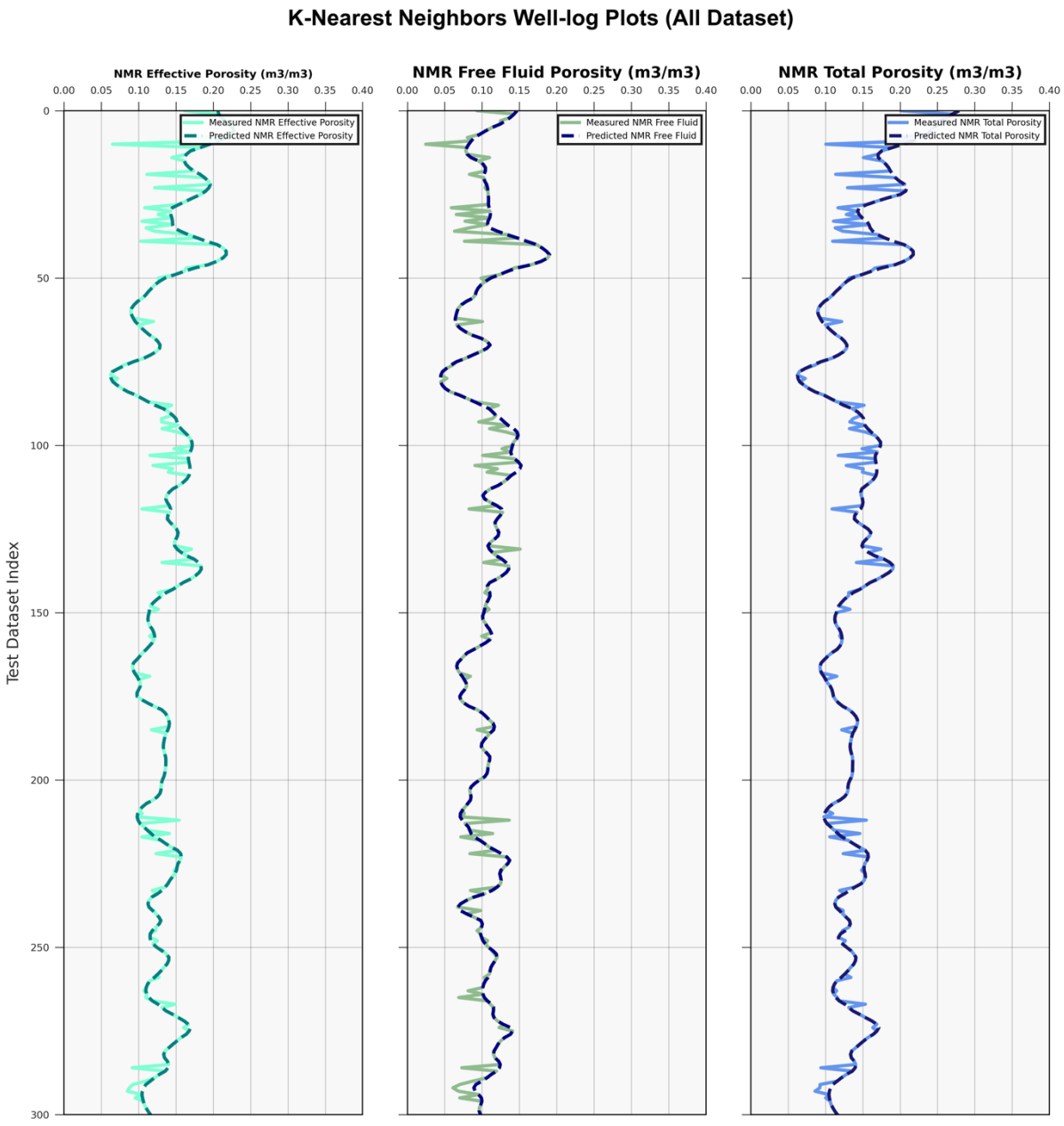

Figure 4-12: Comparing the match between the predicted and measured NMR porosity, for the K-NN Model, across the Entire dataset index of the original dataset contain whole five wells. Left track: Predicted and Measured NMR Effective Porosity (m3/m3). Middle track: Predicted and Measured NMR Free Fluid (m3/m3). Right track: Predicted and Measured NMR Total Porosity (m3/m3).

The observed higher deviation in predicted porosities from measured values in the K-Nearest Neighbors (K-NN) model compared to Random Forest (RF) and Gradient Boosting (GB) models suggests that K-NN might struggle to capture the intricate relationships present in the porosity data. K-NN relies heavily on the proximity of data points, potentially making it more susceptible to noise or outliers, which could contribute to larger prediction errors. On the other hand, Random Forest and Gradient Boosting models might have better accommodated the complexities within the dataset, leading to lower deviations between predicted and measured porosities.

4.1.3 Evaluation of different model performances using the different dataset [1-BRSA-1116-RJS] (Not included in training-test)

4.1.3.1 Scatter plots

When using a model with a dataset it hasn't seen before, lower accuracy is common. The model might struggle to generalize to new data points because it hasn't learned patterns specific to that new dataset. The scatter plots reflecting this mismatch between predicted and measured values often indicate a lack of a clear trend line, showing the disparities between predicted and actual values. The scatter plots of predicted and measured porosities can be found in figures 4-13, 4-14, and 4-15.

This discrepancy could occur due to various reasons:

- **Lack of Generalization:** The model might have learned specific patterns or nuances from the training data that don't hold true in the new dataset. Consequently, its predictions might not align well with this unseen data.
- **Dataset Differences:** The new dataset could have different distributions, ranges, or outliers compared to the training data. These disparities can significantly impact the model's ability to make accurate predictions.
- **Feature Variability:** Features in the new dataset might possess characteristics that the model wasn't trained to handle. It might require additional feature engineering or adjustments to the model architecture to effectively learn these new patterns.
- **Overfitting or Underfitting:** The model might suffer from overfitting (being too specific to the training data) or underfitting (not capturing enough patterns), affecting its performance on new data.

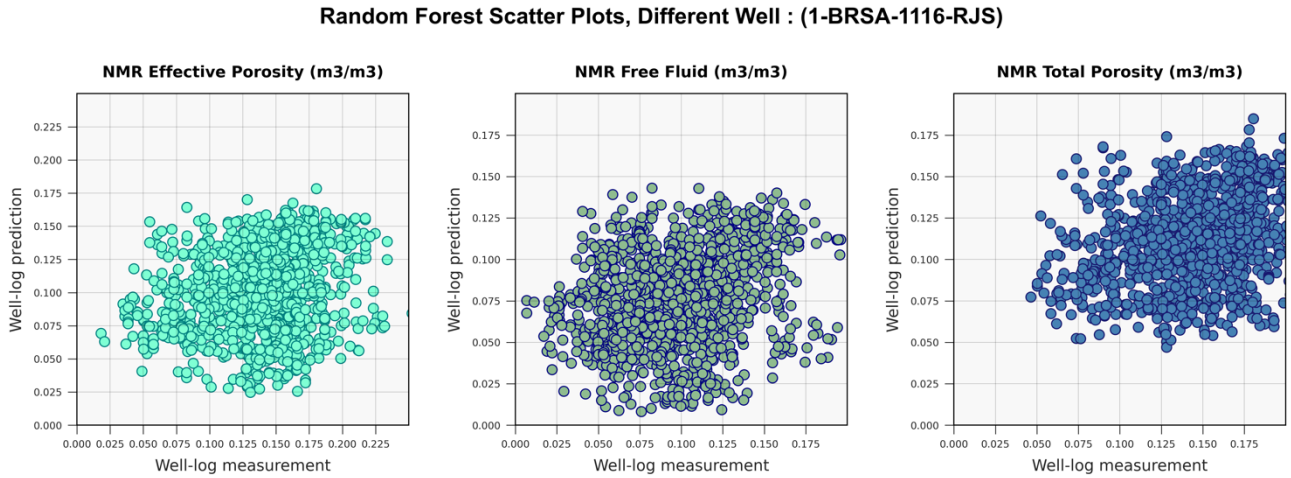

Figure 4-13: Scatter plots of predicted versus measured NMR porosity, for the RF Model, applied to the different well (1-BRSA-1116-RJS) not included in training phase: Left plot: NMR Effective Porosity (m3/m3). Middle plot: NMR Free Fluid (m3/m3). Right plot: NMR Total Porosity (m3/m3)

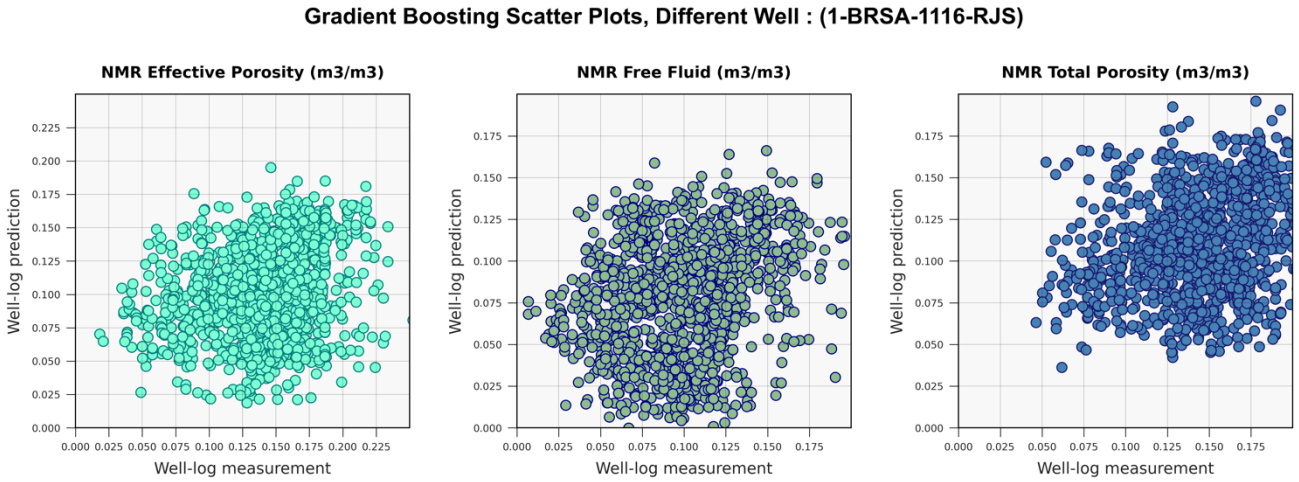

Figure 4-14: Scatter plots of predicted versus measured NMR porosity, for the GB Model, applied to the different well (1-BRSA-1116-RJS) not included in training phase: Left plot: NMR Effective Porosity (m3/m3). Middle plot: NMR Free Fluid (m3/m3). Right plot: NMR Total Porosity (m3/m3).

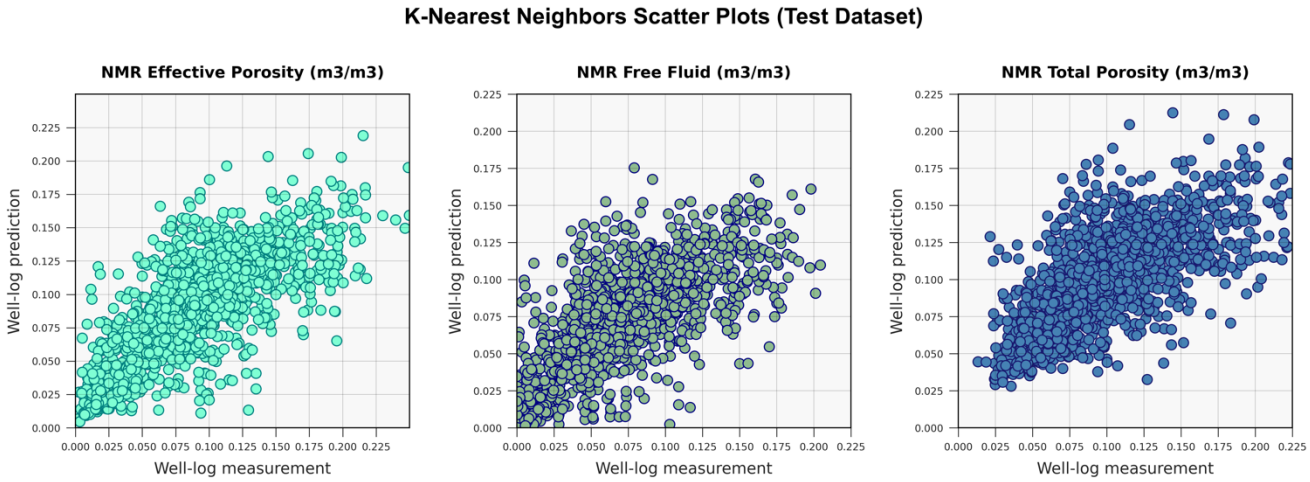

Figure 4-15: Scatter plots of predicted versus measured NMR porosity, for the K-NN Model, applied to the different well (1-BRSA-1116-RJS) not included in training phase: Left plot: NMR Effective Porosity (m3/m3). Middle plot: NMR Free Fluid (m3/m3). Right plot: NMR Total Porosity (m3/m3).

To address this issue, several steps can be taken:

- More Diverse Training Data: Incorporate more diverse and representative data during the model training phase to improve its ability to generalize to new datasets.
- Fine-tuning or Retraining: Retrain the model using the new dataset or fine-tune it by using transfer learning techniques to adapt it to the characteristics of the new data.
- Feature Engineering: Modify or engineer features to better capture the nuances of the new dataset, ensuring the model can learn relevant patterns.
- Regularization Techniques: Apply regularization methods to prevent overfitting or underfitting, enabling the model to generalize better to unseen data.

4.1.3.2 Well-log Plots

In Figure , while examining the Free Fluid, Effective, and Total porosities, it becomes evident that despite discrepancies existing within each metric, there are specific segments where the model's predictions notably align better with the measured data. These instances of improved alignment between predicted and measured values are observable across all three porosity types. The plots of measured and predicted porosities can be found in figures 4-16, 4-17, and 4-18.

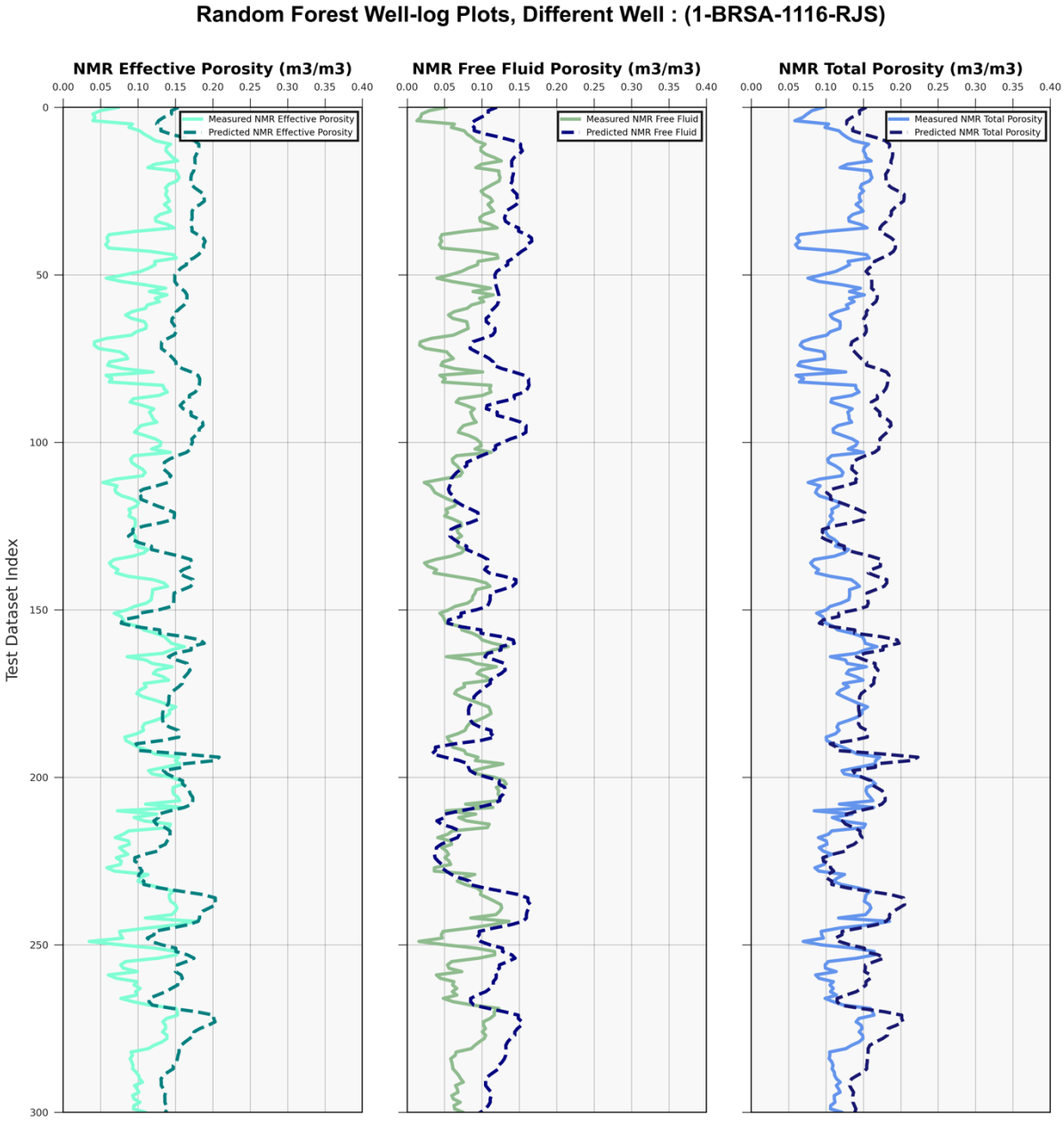

Figure 4-16: Comparing the match between the predicted and measured NMR porosity, for the RF Model, applied to the different well (1-BRSA-1116-RJS): Left track: Predicted and Measured NMR Effective Porosity (m3/m3). Middle track: Predicted and Measured NMR Free Fluid (m3/m3). Right track: Predicted and Measured NMR Total Porosity (m3/m3).

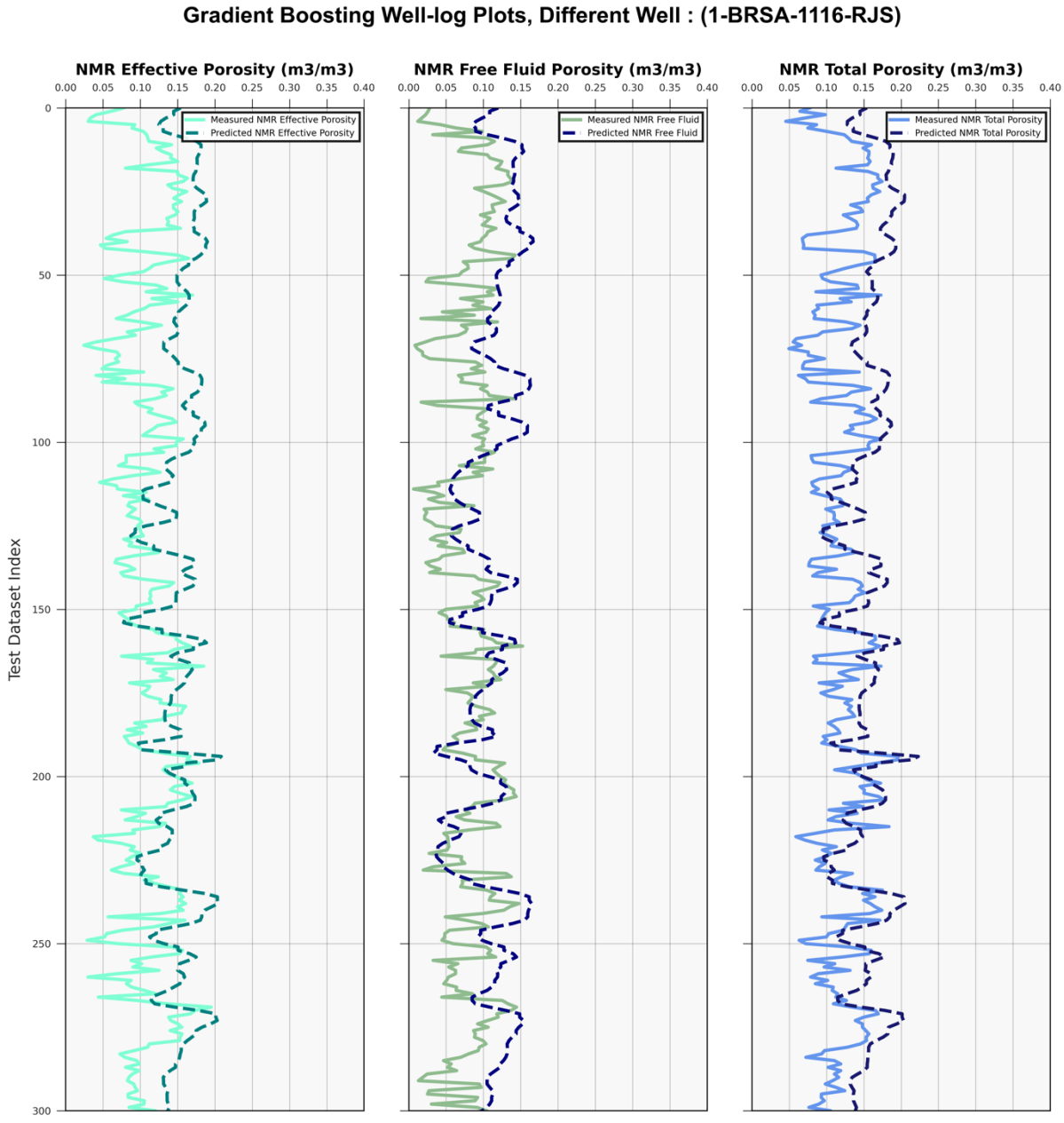

Figure 4-17: Comparing the match between the predicted and measured NMR porosity, for the GB Model, applied to the different well (1-BRSA-1116-RJS): Left track: Predicted and Measured NMR Effective Porosity (m3/m3). Middle track: Predicted and Measured NMR Free Fluid (m3/m3). Right track: Predicted and Measured NMR Total Porosity (m3/m3).

Although discrepancies persist within Free Fluid, Effective, and Total porosity predictions, there are identifiable areas where the model demonstrates a closer match with the actual measured values. These segments might represent specific conditions or characteristics within the well that the model accurately captures, indicating localized instances of enhanced predictive accuracy across all porosity metrics. Identifying these areas of improved alignment could offer valuable insights into the factors

influencing the model's performance and highlight regions where it excels in predicting porosity characteristics within the 1-BRSA-1116-RJS well.

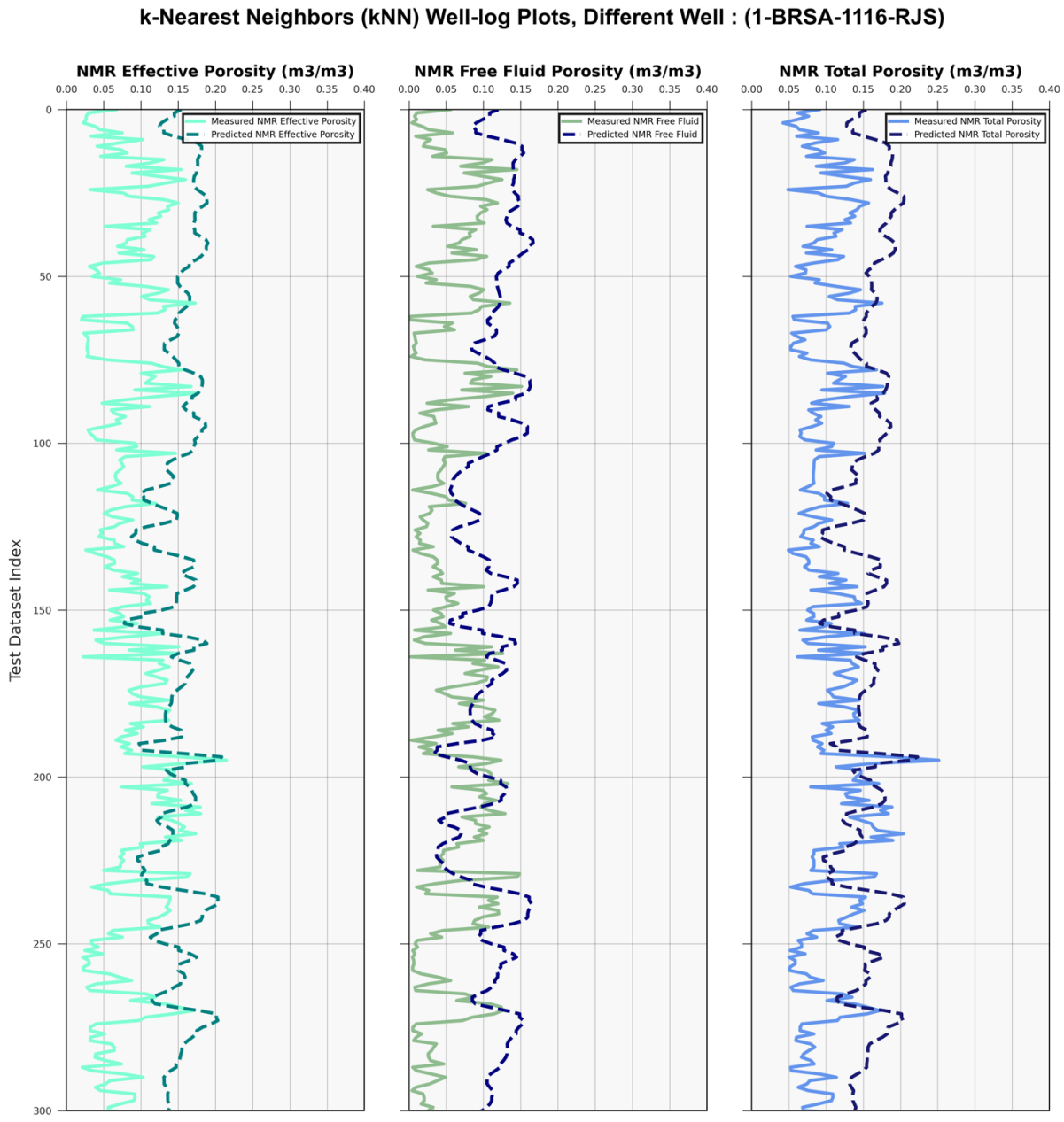

Figure 4-18: Comparing the match between the predicted and measured NMR porosity, for the K-NN Model, applied to the different well (1-BRSA-1116-RJS): Left track: Predicted and Measured NMR Effective Porosity (m3/m3). Middle track: Predicted and Measured NMR Free Fluid (m3/m3). Right track: Predicted and Measured NMR Total Porosity (m3/m3).

4.1.4 Evaluation of different model performances using single well [3-BRSA-1215-RJS] (Included in training-test)

4.1.4.1 Scatter plots

Using the well 3-BRSA-1215-RJS in the training set is a deliberate choice aimed at strengthening my model. This well stands out due to its unique formation, differing from others in the dataset. Its inclusion during training allows the model to adapt to this distinct geological aspect, enhancing its ability to handle varied formations and potentially improving its overall predictive capability. The scatter plots of predicted and measured porosities can be found in figures 4-19, 4-20, and 4-21.

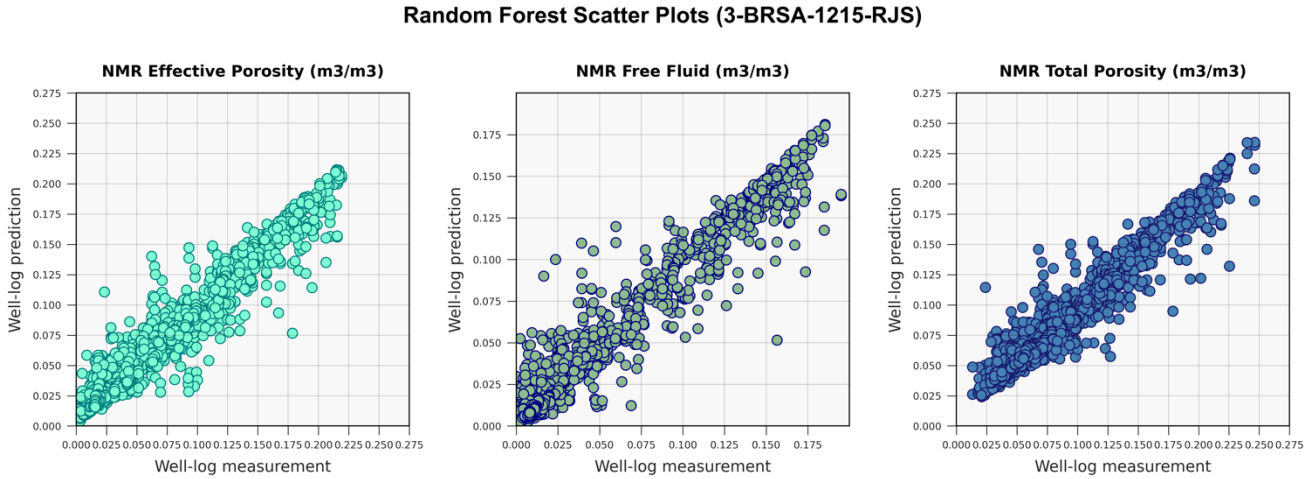

Figure 4-19: Scatter plots of predicted versus measured NMR porosity, for the RF Model, applied to one of the well (3-BRSA-1215-RJS) included in training phase: Left plot: NMR Effective Porosity (m3/m3). Middle plot: NMR Free Fluid (m3/m3). Right plot: NMR Total Porosity (m3/m3).

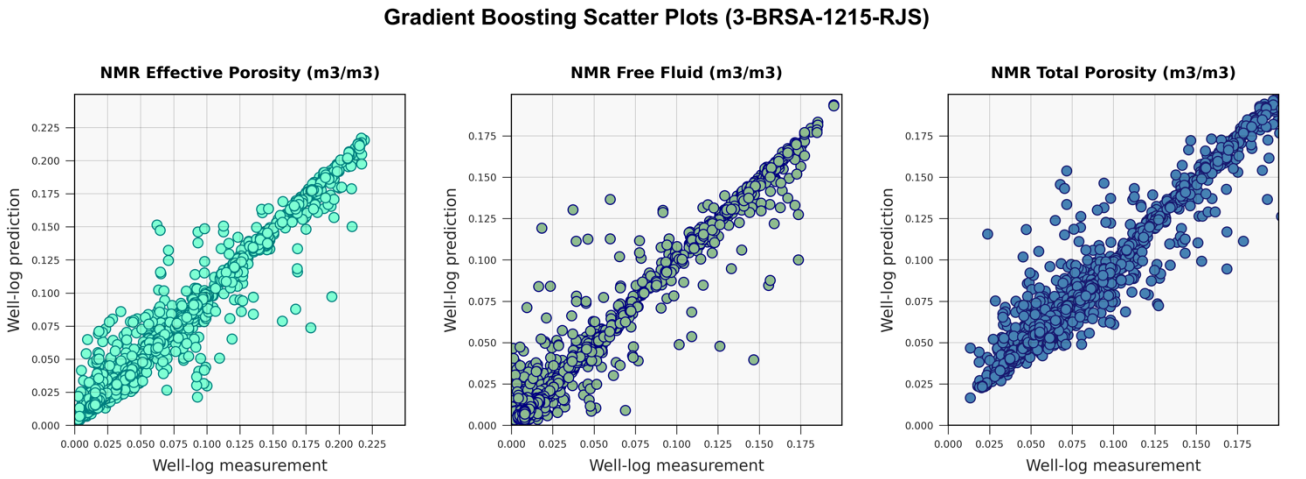

Figure 4-20: Scatter plots of predicted versus measured NMR porosity, for the GB Model, applied to one of the well (3-BRSA-1215-RJS) included in training phase: Left plot: NMR Effective Porosity (m3/m3). Middle plot: NMR Free Fluid (m3/m3). Right plot: NMR Total Porosity (m3/m3).

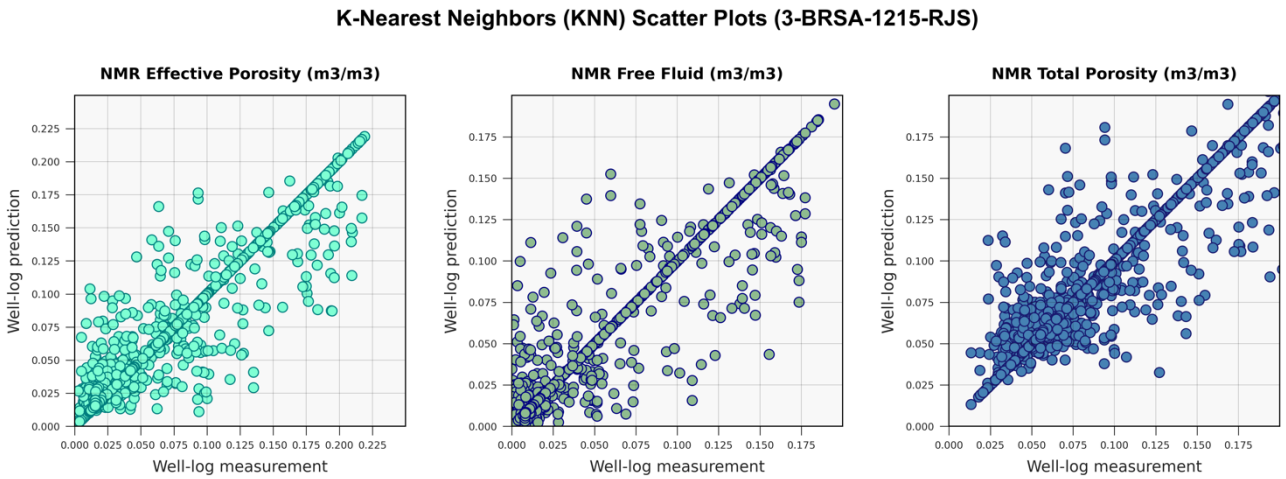

Figure 4-21: Scatter plots of predicted versus measured NMR porosity, for the K-NN Model, applied to one of the well (3-BRSA-1215-RJS) included in training phase: Left plot: NMR Effective Porosity (m3/m3). Middle plot: NMR Free Fluid (m3/m3). Right plot: NMR Total Porosity (m3/m3).

4.1.4.2 Well-log Plots

The comprehensive assessment of the Random Forest, Gradient Boosting, and K-nearest Neighbour model's predictive capacity for key petrophysical parameters—NMR porosity (Effective, Free fluid, Total), as well as Predicted Permeability and Saturation Using the Timur-Coates Correlation—offers invaluable insights into its efficacy. Analyzing the match between predicted and measured NMR porosities across multiple tracks illuminates the model's performance. The first three tracks portray the comparison between predicted and measured NMR Effective, Free Fluid, and Total Porosities (measured in m^3/ m^3). These tracks highlight the model's ability to capture the nuances of porosity distribution within the well (3-BRSA-1215-RJS) utilized within the training set. The fourth track extends this evaluation by putting together the model's predictions of Effective, Free Fluid, and Total Porosities, providing a holistic view of its overall estimation capabilities. Moreover, the subsequent tracks, the fifth and sixth ones, exhibit the model's forecasted values for permeability (mD) and irreducible water saturation (expressed as a percentage) based on the Timur-Coates Correlation. The plots of predicted and measured porosities and Peameability can be found in figures 4-22, 4-23, and 4-24.

Random Forest:

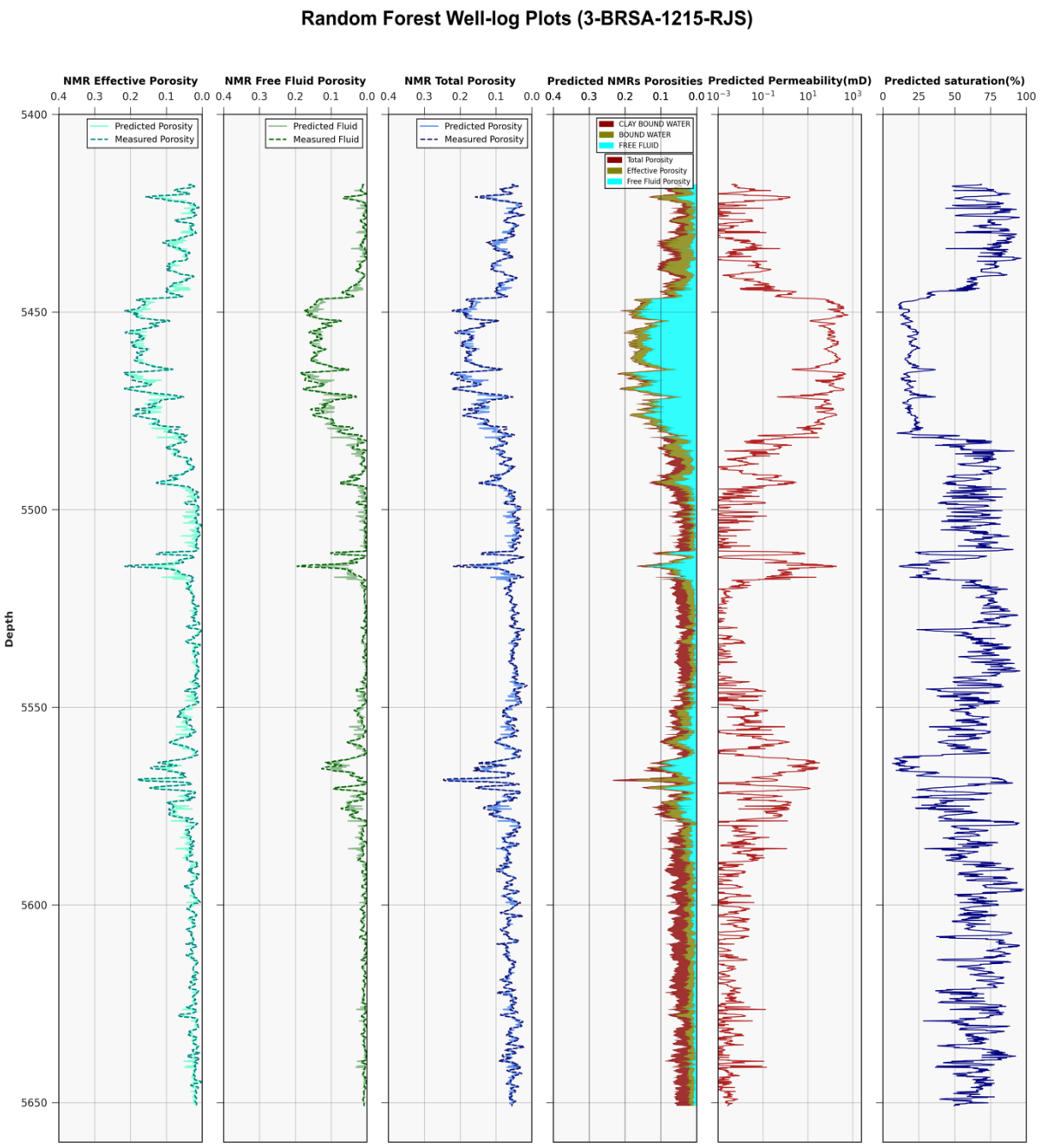

Figure 4-22: Comparing the match between the predicted and measured NMR porosity (Effective, Free fluid Total), and Predicted Permeability & Saturation Using Timur-Coates Correlation, for the RF Model, applied to the well (3-BRSA-1215-RJS) included in training phase: (From left to right) 1st track: Predicted and Measured NMR Effective Porosity (m3/m3). 2nd track: Predicted and Measured NMR Free Fluid (m3/m3). 3rd track: Predicted and Measured NMR Total Porosity (m3/m3). 4th track: Predicted Effective, Free Fluid, Total Porosities (m3/m3), 5th track: Timur-Coates Predicted Permeability (mD), 6th track: Timur-Coates Predicted Irreducible water Saturation (%)

Gradient Boosting:

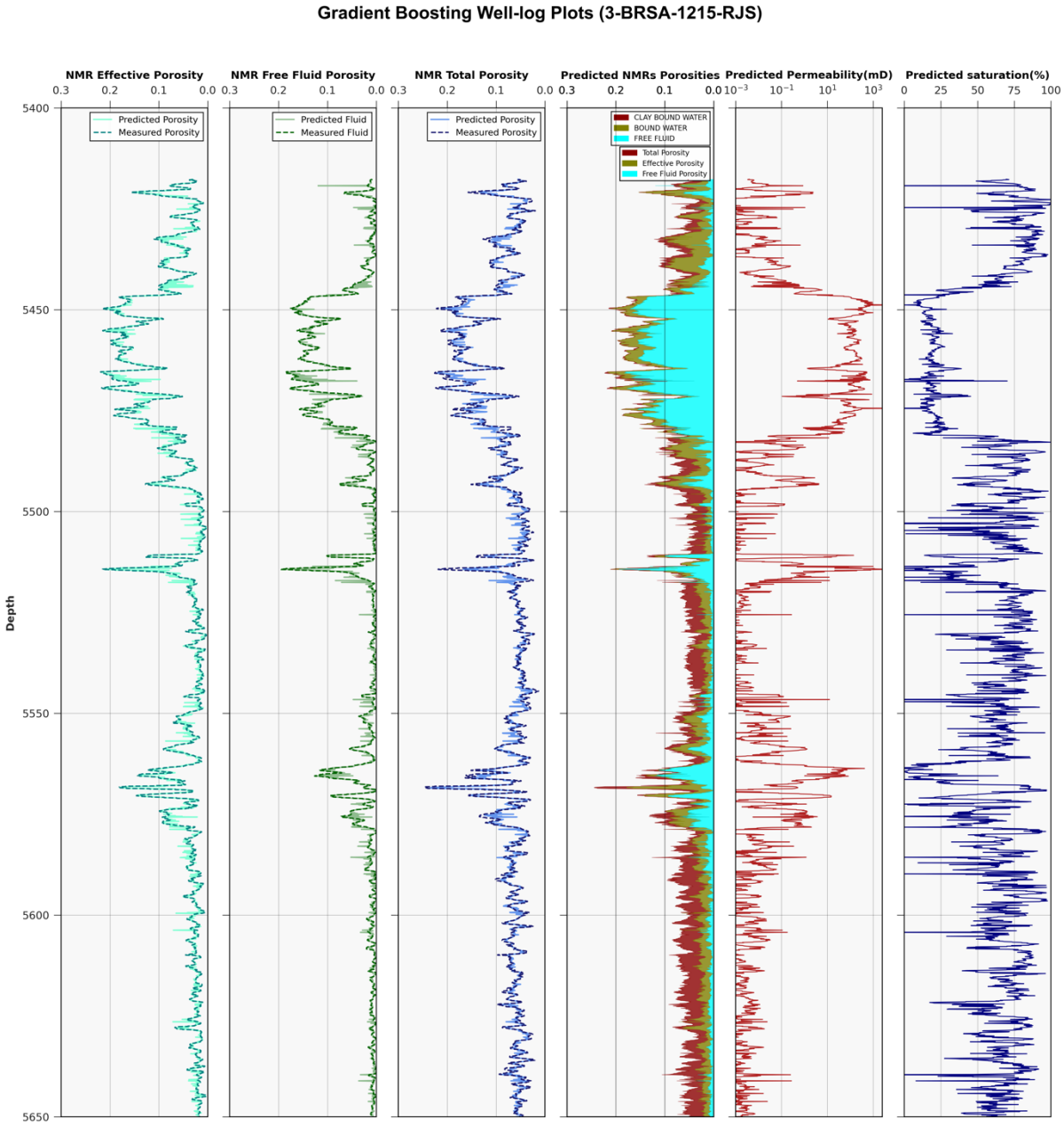

Figure 4-23: Comparing the match between the predicted and measured NMR porosity (Effective, Free fluid Total), and Predicted Permeability & Saturation Using Timur-Coates Correlation, for the GB Model, applied to the well (3-BRSA-1215-RJS) included in training phase: (From left to right) 1st track: Predicted and Measured NMR Effective Porosity (m3/m3). 2nd track: Predicted and Measured NMR Free Fluid (m3/m3). 3rd track: Predicted and Measured NMR Total Porosity (m3/m3). 4th track: Predicted Effective, Free Fluid, Total Porosities (m3/m3), 5th track: Timur-Coates Predicted Permeability (mD), 6th track: Timur-Coates Predicted Irreducible water Saturation (%)

K-Nearest Neighbor (KNN):

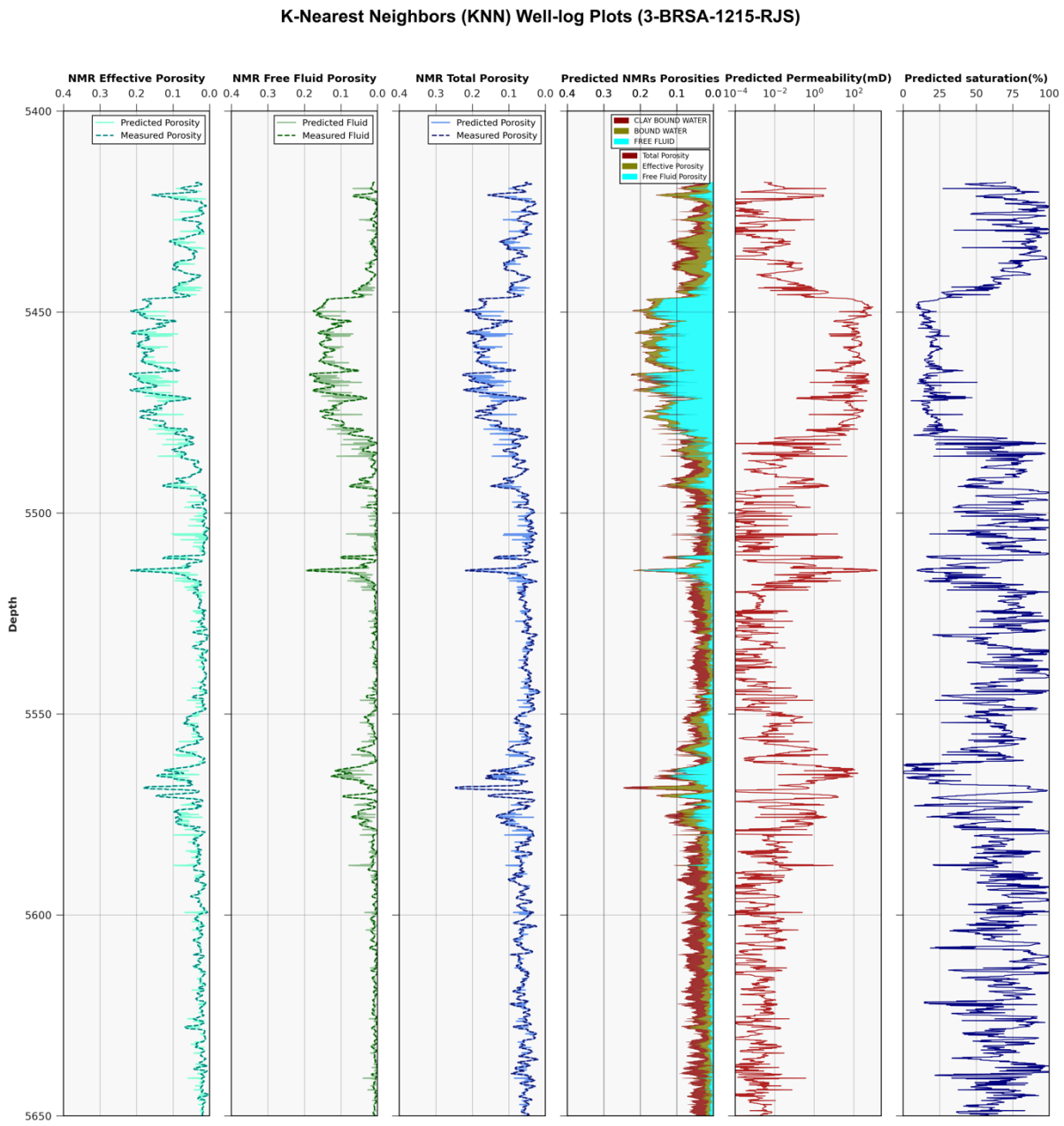

Figure 4-24: Comparing the match between the predicted and measured NMR porosity (Effective, Free fluid Total), and Predicted Permeability & Saturation Using Timur-Coates Correlation, for the K-NN Model, applied to the well (3-BRSA-1215-RJS) included in training phase: (From left to right) 1st track: Predicted and Measured NMR Effective Porosity (m3/m3). 2nd track: Predicted and Measured NMR Free Fluid (m3/m3). 3rd track: Predicted and Measured NMR Total Porosity (m3/m3). 4th track: Predicted Effective, Free Fluid, Total Porosities (m3/m3), 5th track: Timur-Coates Predicted Permeability (mD), 6th track: Timur-Coates Predicted Irreducible water Saturation (%)

4.1.4.3 Well-log and Lab data plots

The comprehensive assessment across the Random Forest, Gradient Boosting, and K-nearest Neighbour models on the well (3-BRSA-1215-RJS) brings to light their predictive abilities for permeability and porosity against laboratory measurements. Through the initial tracks, the models' proximity to laboratory-measured values for permeability (mD) and porosity percentages demonstrates their adeptness in estimating fluid flow characteristics and available pore space within the rock. The inclusion of a scatter plot in the third track enriches this evaluation, visually representing the relationship between permeability and porosity and allowing insight into how well the models capture intrinsic correlations seen in laboratory data. Furthermore, the subsequent histograms depicting the distributions of porosity and permeability values, both predicted and laboratory-based, facilitate a comparative analysis of their patterns. This comprehensive assessment delves not only into the alignment between predicted and laboratory values but also scrutinizes the models' abilities to replicate trends and distribution patterns. Ultimately, this multifaceted evaluation provides a comprehensive understanding of the models' reliability in estimating these fundamental rock properties.

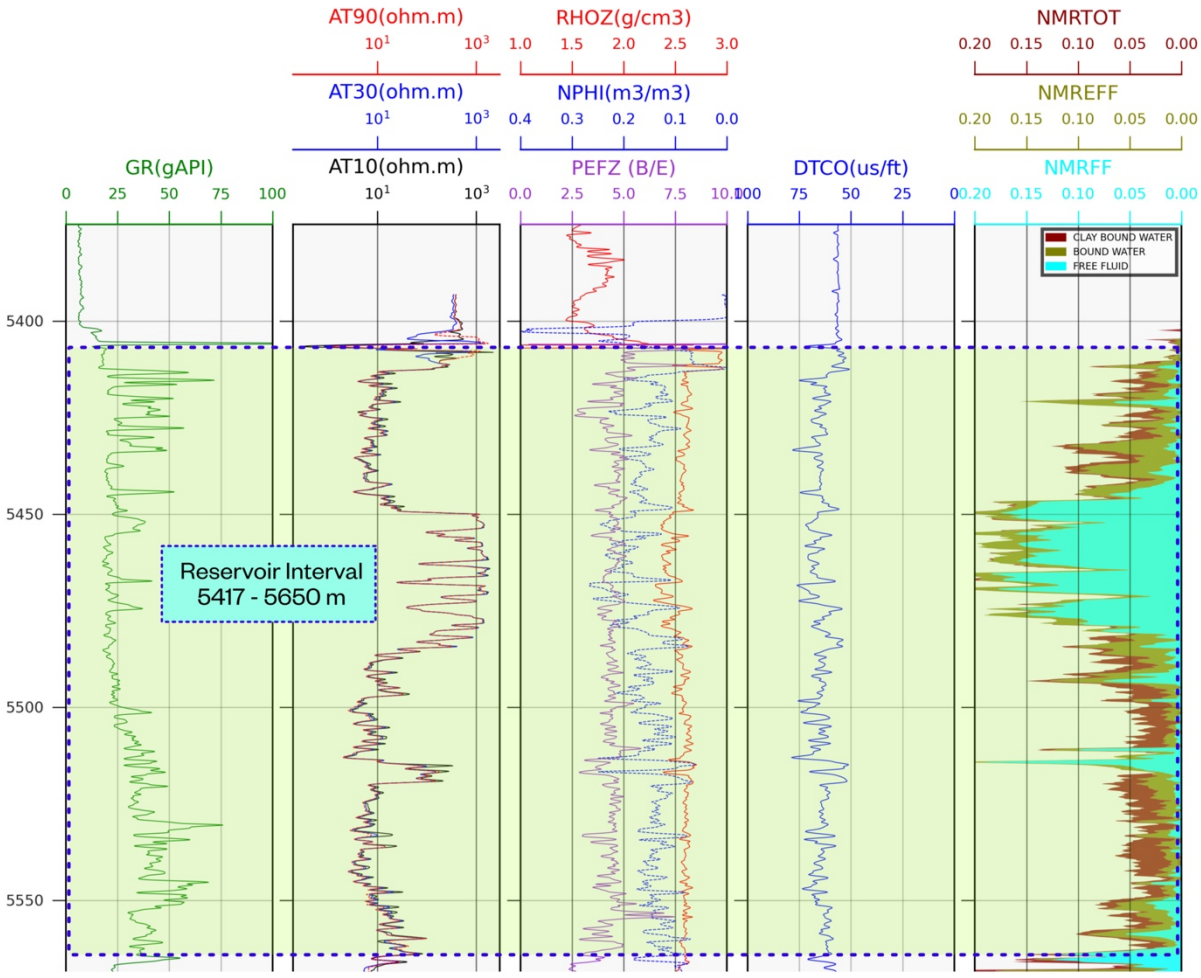

Figure 4-25: The reservoir interval of the well 3-BRSA-1215-RJS

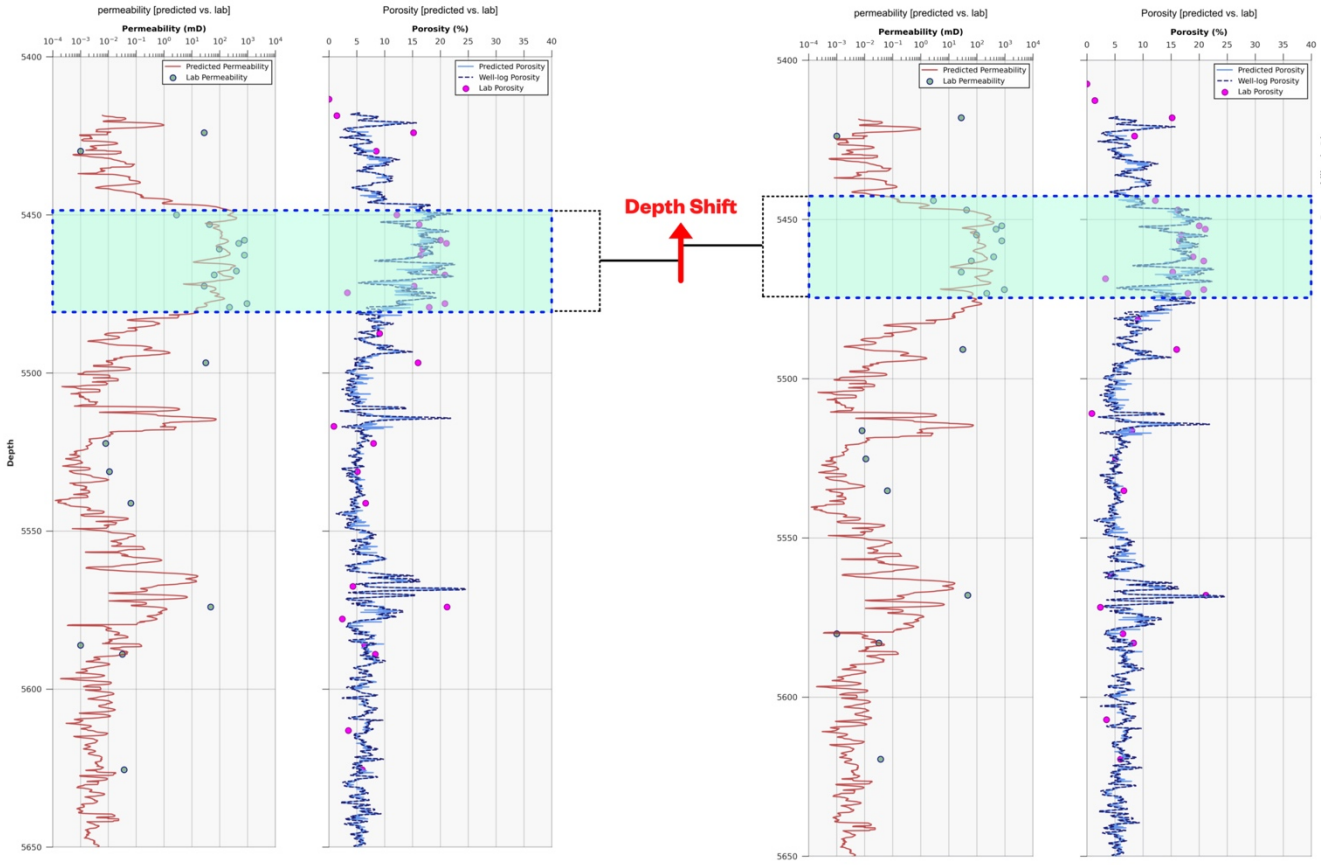

Figure 4-26: Investigation of depth shift on log data in the well 3-BRSA-1215-RJS

It is quite common to be some discrepancies between predicted results from logging operations and actual laboratory measurements. This is because there are a number of factors that can affect the accuracy of logging data, including cable length, logging standards, and the type of rock being logged.

In the case of well 3-BRSA-1215-RJS, it appears that the depth shift on the lab data has helped to improve the coherency between predicted and actual results. This suggests that the original logging data may have been inaccurate due to the differences in cable length and logging standards between the logging operation and the laboratory measurement.

The plots of predicted and measured porosities and permeability versus laboratory data can be found in figures 4-27, 4-28, and 4-29.

Random Forest:

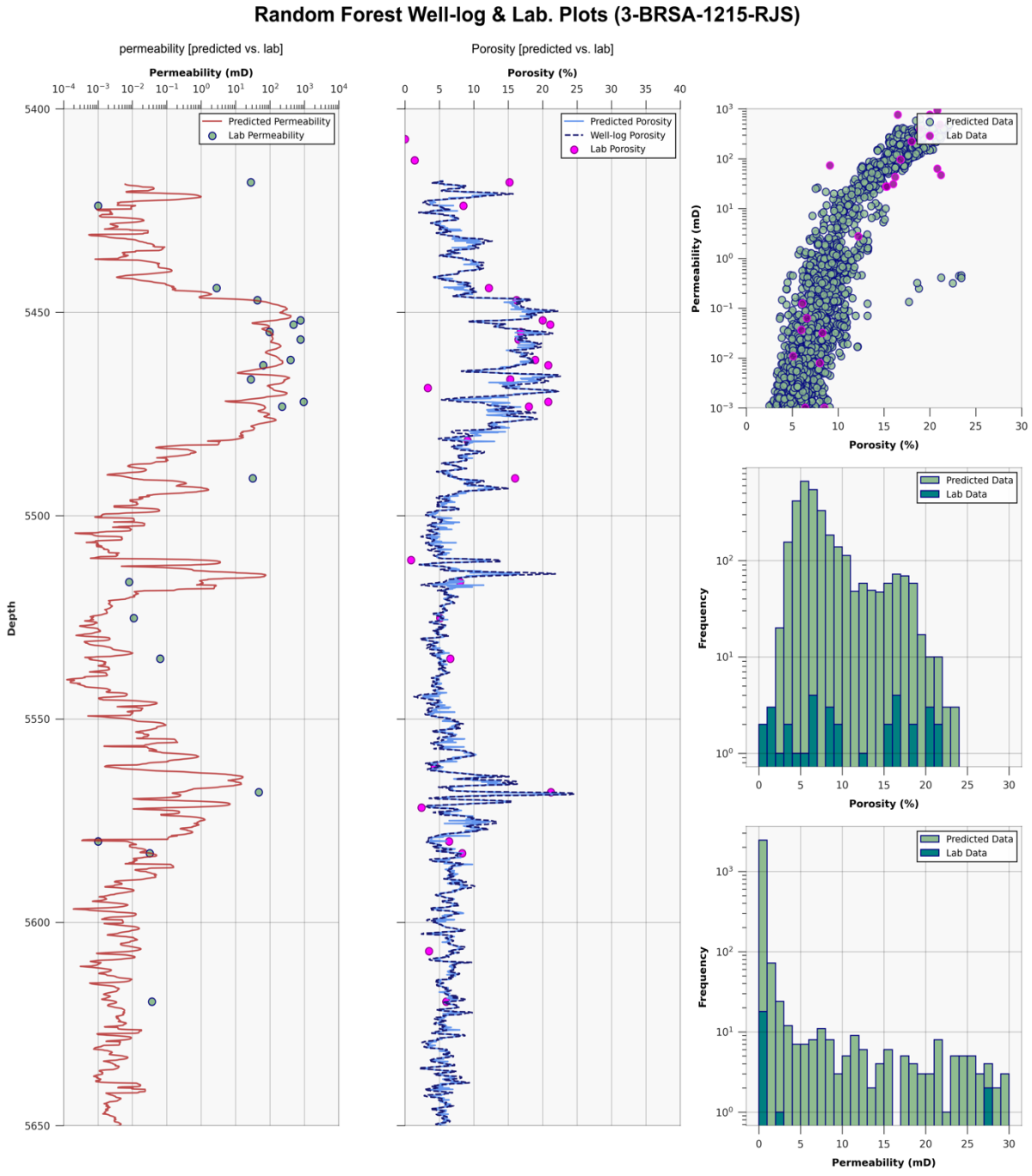

Figure 4-27: Comparing the match between the predicted & laboratory measurement of permeability and porosity, Comparing scatter plot permeability-porosity of predicted and laboratory data, Histogram of Porosity and Permeability distribution of predicted and laboratory data, for the RF Model, applied to the well (3-BRSA-1215-RJS) included in training set (for Permeability Prediction Timur-Coates Correlation has been used): (From left to right) 1st track: Predicted and laboratory Permeability (mD). 2nd track: Predicted and Measured Porosity (%). 3rd track: Scatter plot of permeability-porosity of predicted and laboratory data (mD vs. %), 4th track: Histogram of Porosity of predicted and laboratory data (%), 5th track Histogram of Permeability of predicted and laboratory data (mD)

Gradient Boosting:

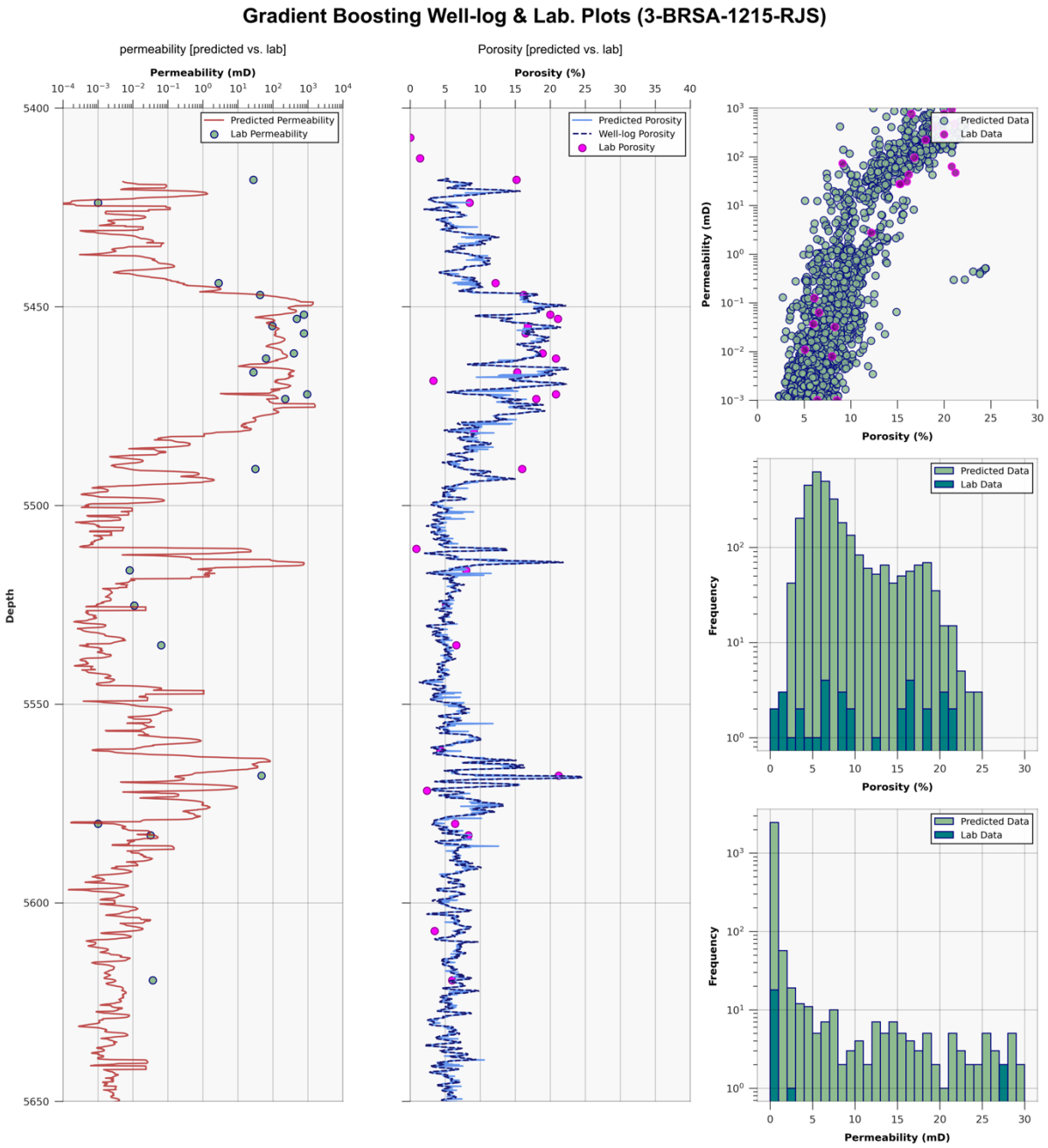

Figure 4-28: Comparing the match between the predicted & laboratory measurement of permeability and porosity, Comparing scatter plot permeability-porosity of predicted and laboratory data, Histogram of Porosity and Permeability distribution of predicted and laboratory data, for the GB Model, applied to the well (3-BRSA-1215-RJS) included in training set (for Permeability Prediction Timur-Coates Correlation has been used): (From left to right) 1st track: Predicted and laboratory Permeability (mD). 2nd track: Predicted and Measured Porosity (%). 3rd track: Scatter plot of permeability-porosity of predicted and laboratory data (mD vs. %), 4th track: Histogram of Porosity of predicted and laboratory data (%), 5th track Histogram of Permeability of predicted and laboratory data (mD)

K-Nearest Neighbor (KNN):

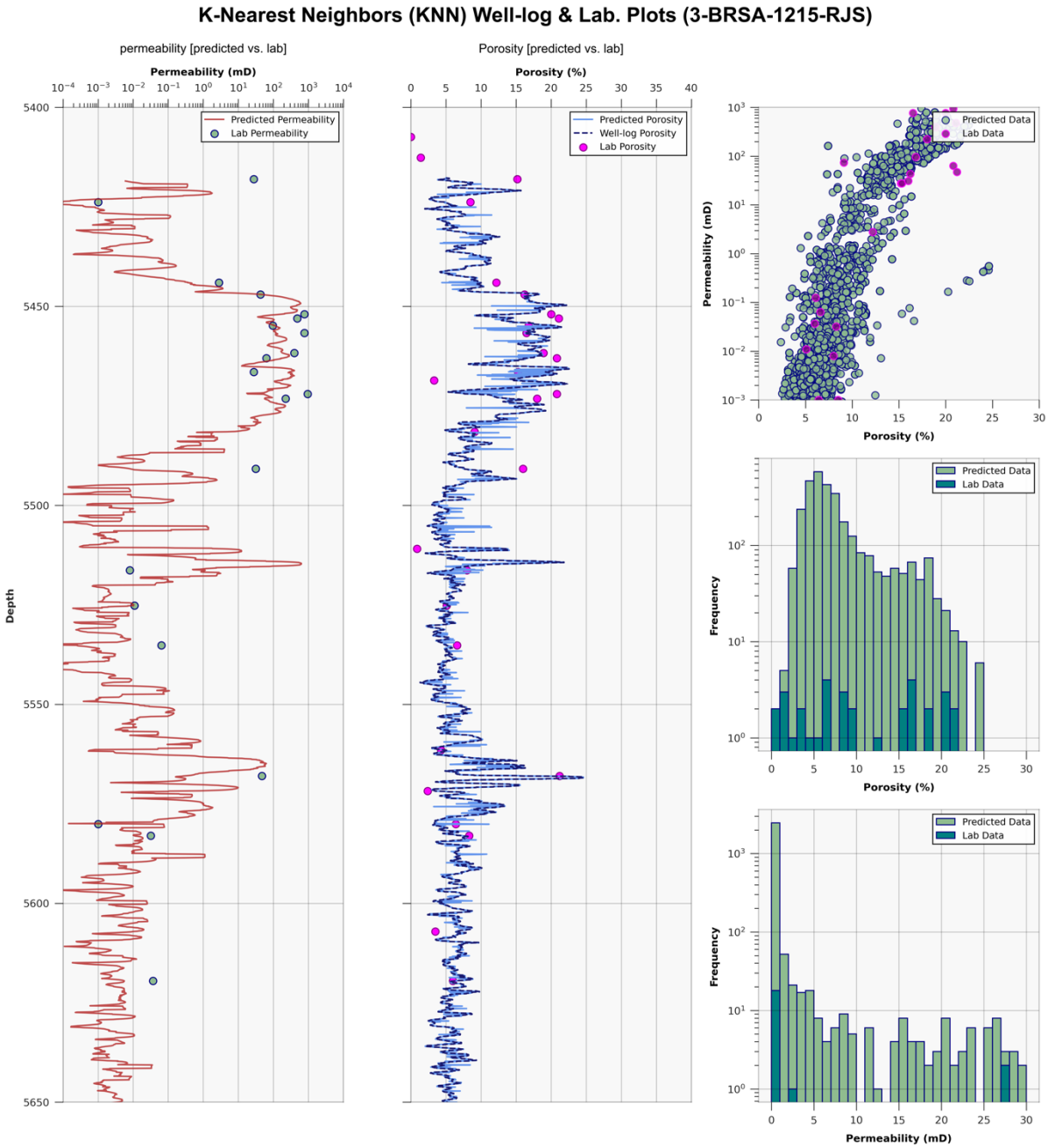

Figure 4-29: Comparing the match between the predicted & laboratory measurement of permeability and porosity, Comparing scatter plot permeability-porosity of predicted and laboratory data, Histogram of Porosity and Permeability distribution of predicted and laboratory data, for the K-NN Model, applied to the well (3-BRSA-1215-RJS) included in training set (for Permeability Prediction Timur-Coates Correlation has been used): (From left to right) 1st track: Predicted and laboratory Permeability (mD). 2nd track: Predicted and Measured Porosity (%). 3rd track: Scatter plot of permeability-porosity of predicted and laboratory data (mD vs. %), 4th track: Histogram of Porosity of predicted and laboratory data (%), 5th track Histogram of Permeability of predicted and laboratory data (mD)

4.1.5 Evaluation of different model performances using single well [1-BRSA-1116-RJS] (Not included in training-test)

4.1.5.1 Well-log Plots

Random Forest:

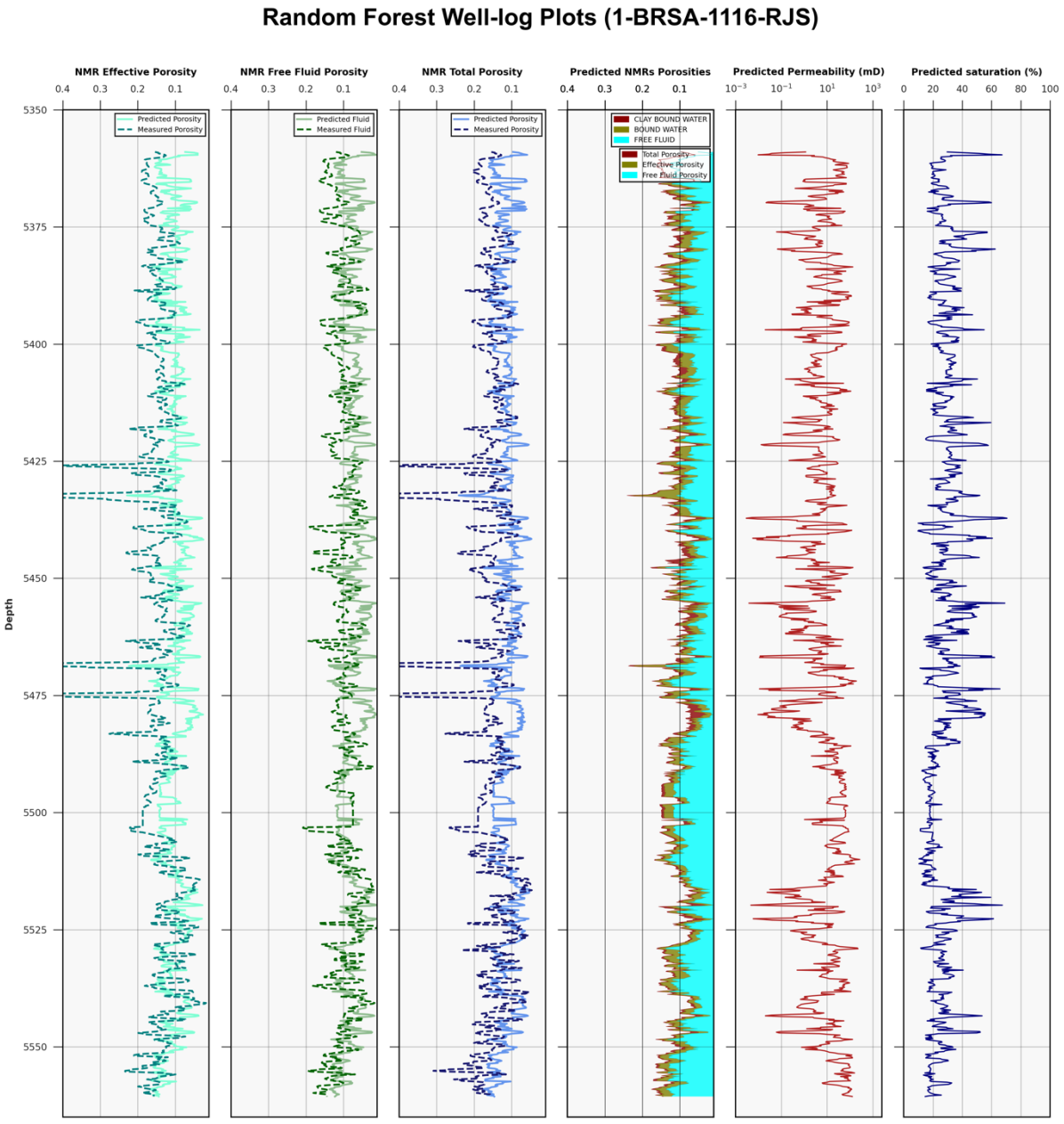

Figure 4-30: Comparing the match between the predicted and measured NMR porosity (Effective, Free fluid Total), and Predicted Permeability & Saturation Using Timur-Coates Correlation, for the RF Model, applied to a different well (3-BRSA-1116-RJS) not included in training phase: (From left to right) 1st track: Predicted and Measured NMR Effective Porosity (m3/m3). 2nd track: Predicted and Measured NMR Free Fluid (m3/m3). 3rd track: Predicted and Measured NMR Total Porosity (m3/m3). 4th track: Predicted Effective, Free Fluid, Total Porosities (m3/m3), 5th track: Timur-Coates Predicted Permeability (mD), 6th track: Timur-Coates Predicted Irreducible water Saturation (%)

Gradient Boosting:

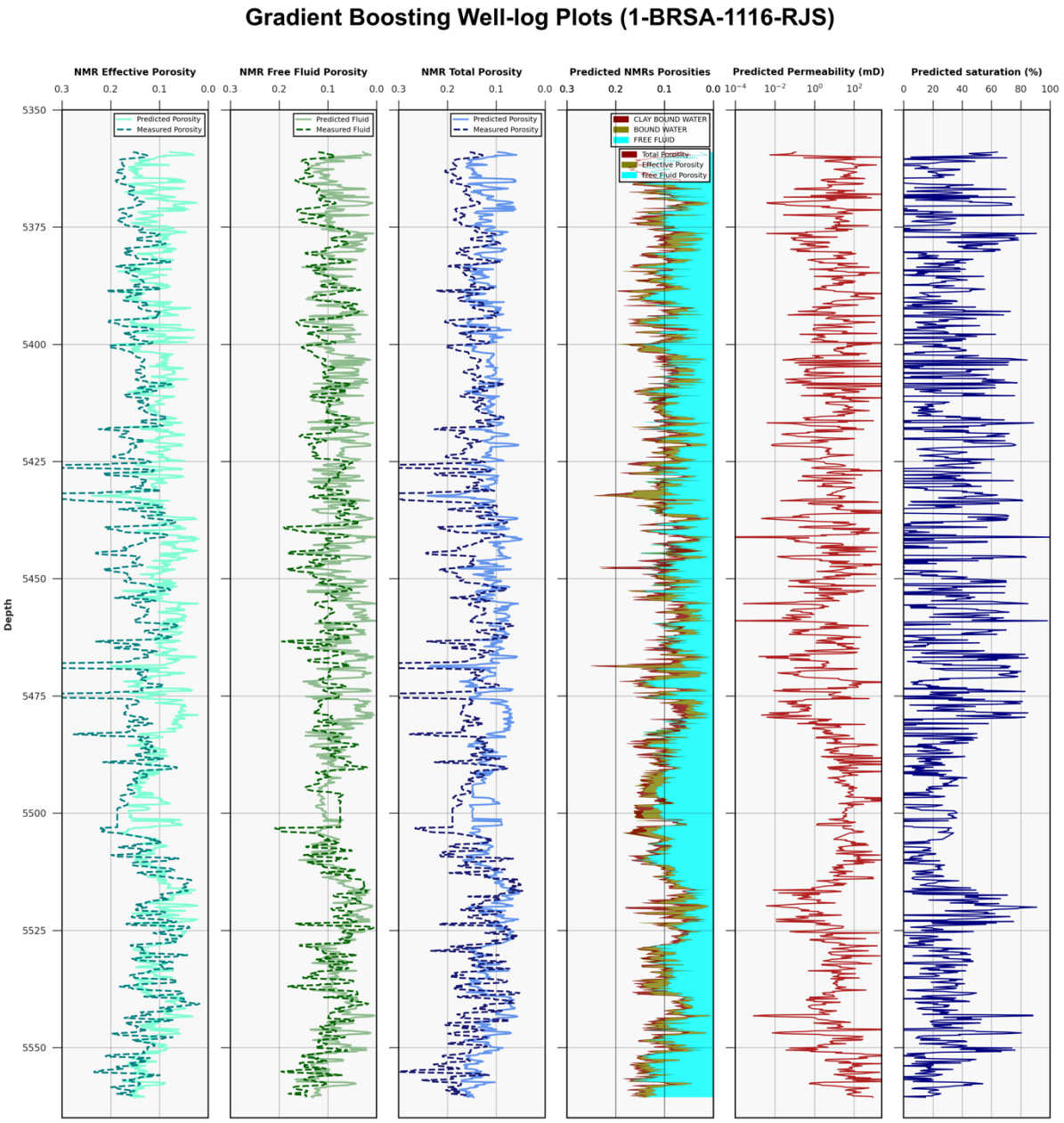

Figure 4-31: Comparing the match between the predicted and measured NMR porosity (Effective, Free fluid Total), and Predicted Permeability & Saturation Using Timur-Coates Correlation, for the GB Model, applied to a different well (3-BRSA-1116-RJS) not included in training phase: (From left to right) 1st track: Predicted and Measured NMR Effective Porosity (m3/m3). 2nd track: Predicted and Measured NMR Free Fluid (m3/m3). 3rd track: Predicted and Measured NMR Total Porosity (m3/m3). 4th track: Predicted Effective, Free Fluid, Total Porosities (m3/m3), 5th track: Timur-Coates Predicted Permeability (mD), 6th track: Timur-Coates Predicted Irreducible water Saturation (%)

K-Nearest Neighbor (KNN):

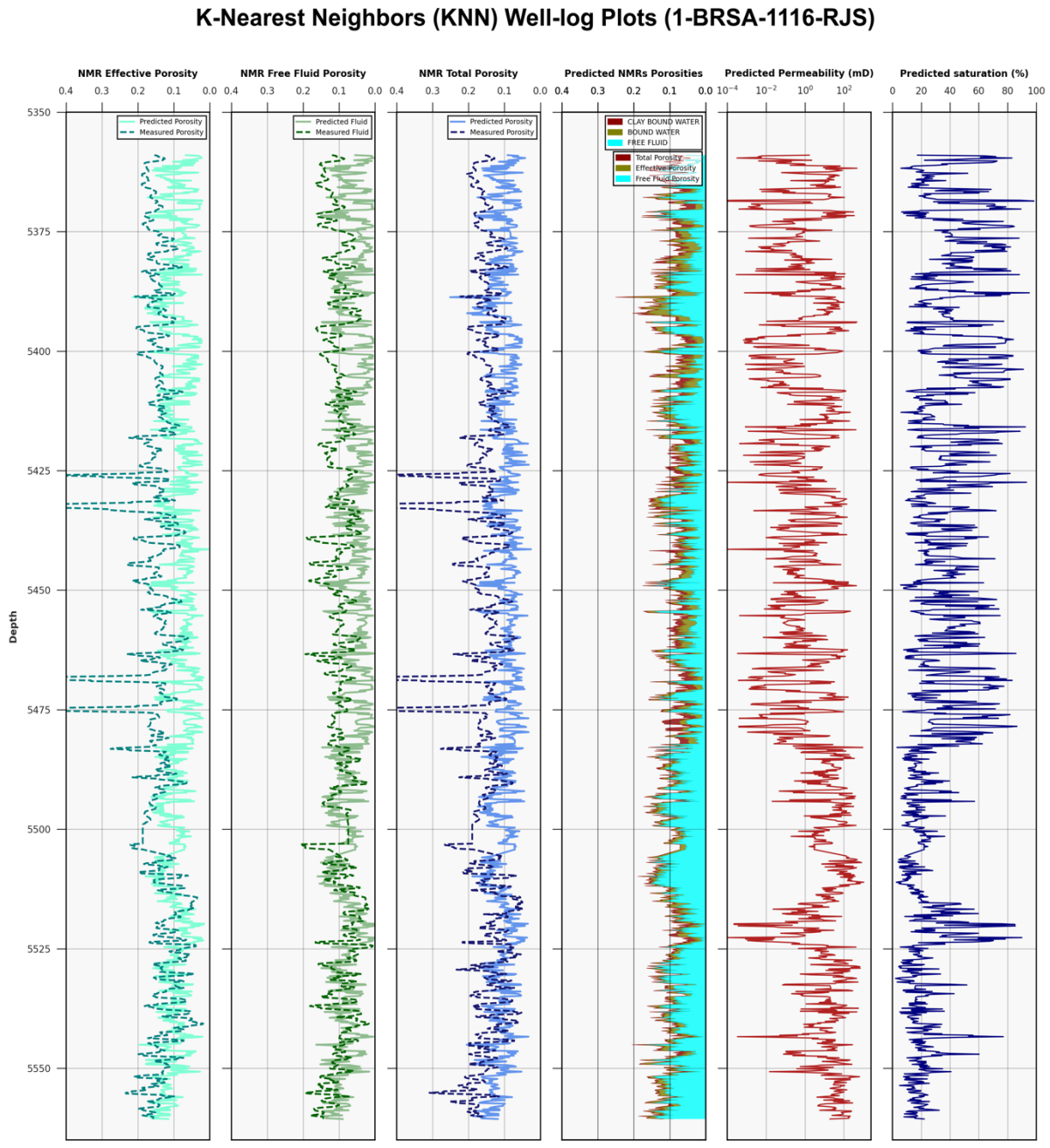

Figure 4-32: Comparing the match between the predicted and measured NMR porosity (Effective, Free fluid Total), and Predicted Permeability & Saturation Using Timur-Coates Correlation, for the K-NN Model, applied to a different well (3-BRSA-1116-RJS) not included in training phase: (From left to right) 1st track: Predicted and Measured NMR Effective Porosity (m3/m3). 2nd track: Predicted and Measured NMR Free Fluid (m3/m3). 3rd track: Predicted and Measured NMR Total Porosity (m3/m3). 4th track: Predicted Effective, Free Fluid, Total Porosities (m3/m3), 5th track: Timur-Coates Predicted Permeability (mD), 6th track: Timur-Coates Predicted Irreducible water Saturation (%)

4.1.5.2 Well-log and Lab data plots

Random Forest:

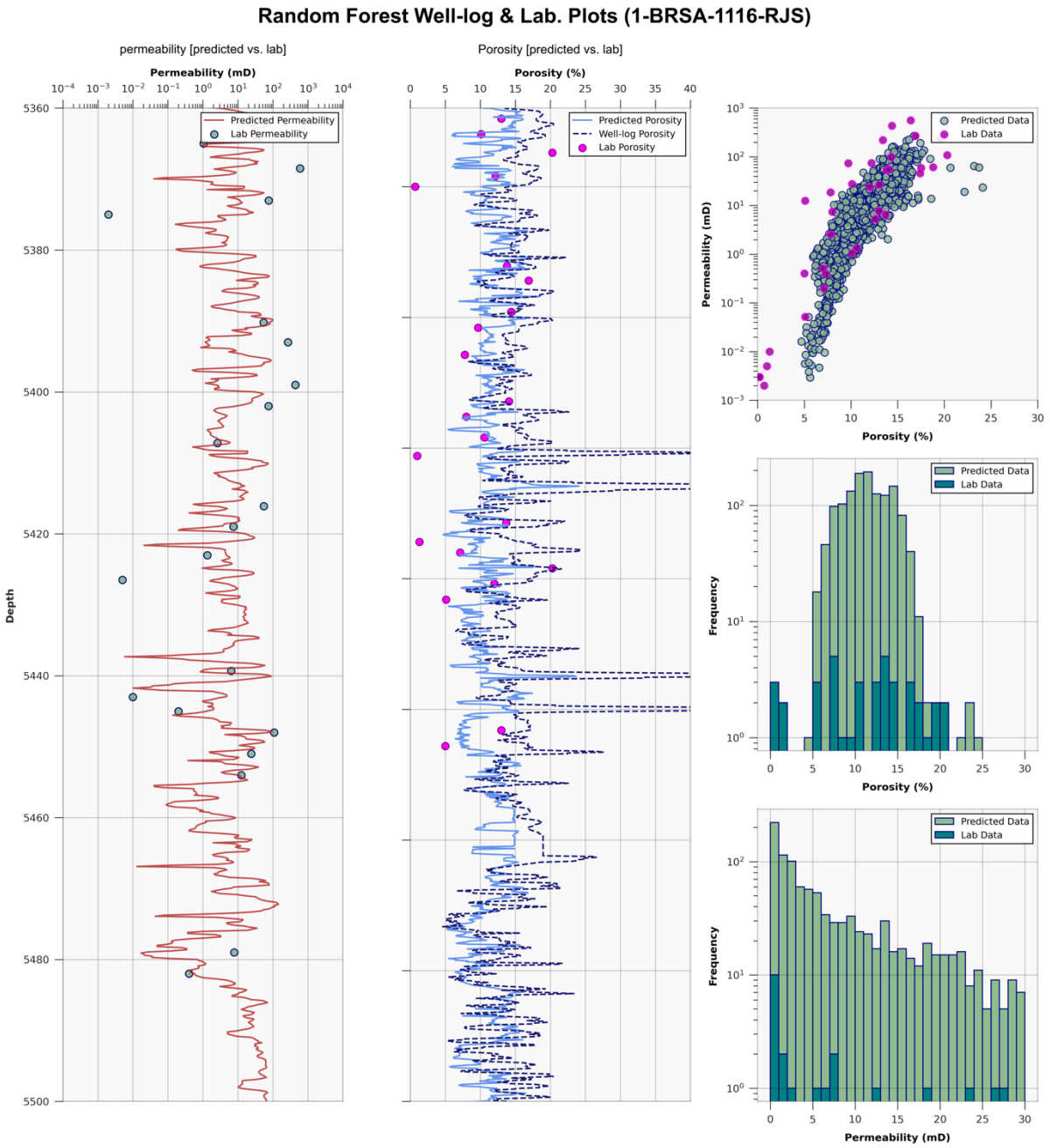

Figure 4-33: Comparing the match between the predicted & laboratory measurement of permeability and porosity, Comparing scatter plot permeability-porosity of predicted and laboratory data, Histogram of Porosity and Permeability distribution of predicted and laboratory data, for the RF Model, applied to a different well (3-BRSA-1116-RJS) not included in training set (for Permeability Prediction Timur-Coates Correlation has been used): (From left to right) 1st track: Predicted and laboratory Permeability (mD). 2nd track: Predicted and Measured Porosity (%). 3rd track: Scatter plot of permeability-porosity of predicted and laboratory data (mD vs. %), 4th track: Histogram of Porosity of predicted and laboratory data (%), 5th track Histogram of Permeability of predicted and laboratory data (mD)

Gradient Boosting:

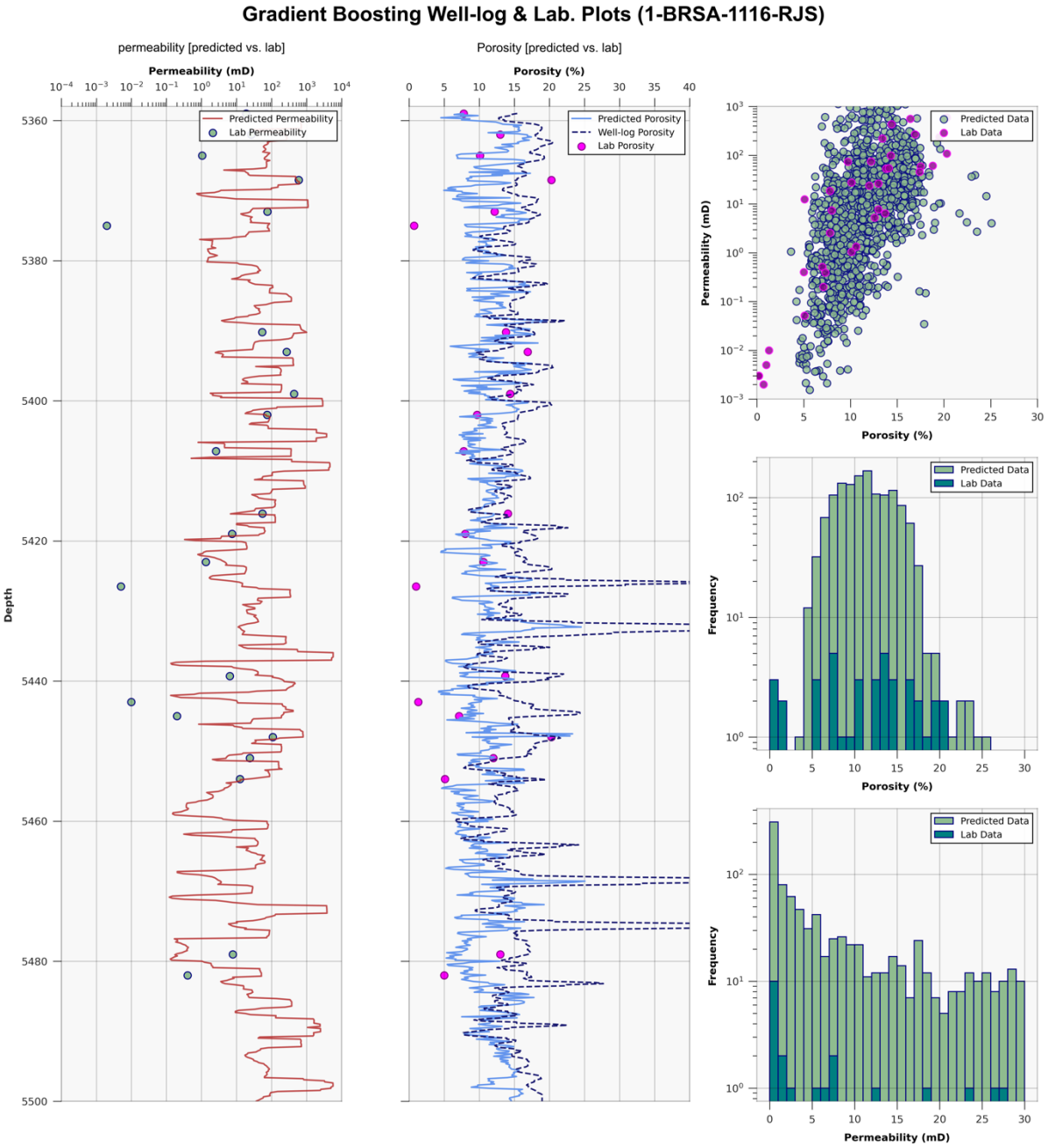

Figure 4-34: Comparing the match between the predicted & laboratory measurement of permeability and porosity, Comparing scatter plot permeability-porosity of predicted and laboratory data, Histogram of Porosity and Permeability distribution of predicted and laboratory data, for the GB Model, applied to a different well (3-BRSA-1116-RJS) not included in training set (for Permeability Prediction Timur-Coates Correlation has been used): (From left to right) 1st track: Predicted and laboratory Permeability (mD). 2nd track: Predicted and Measured Porosity (%). 3rd track: Scatter plot of permeability-porosity of predicted and laboratory data (mD vs. %), 4th track: Histogram of Porosity of predicted and laboratory data (%), 5th track Histogram of Permeability of predicted and laboratory data (mD)

K-Nearest Neighbor (KNN):

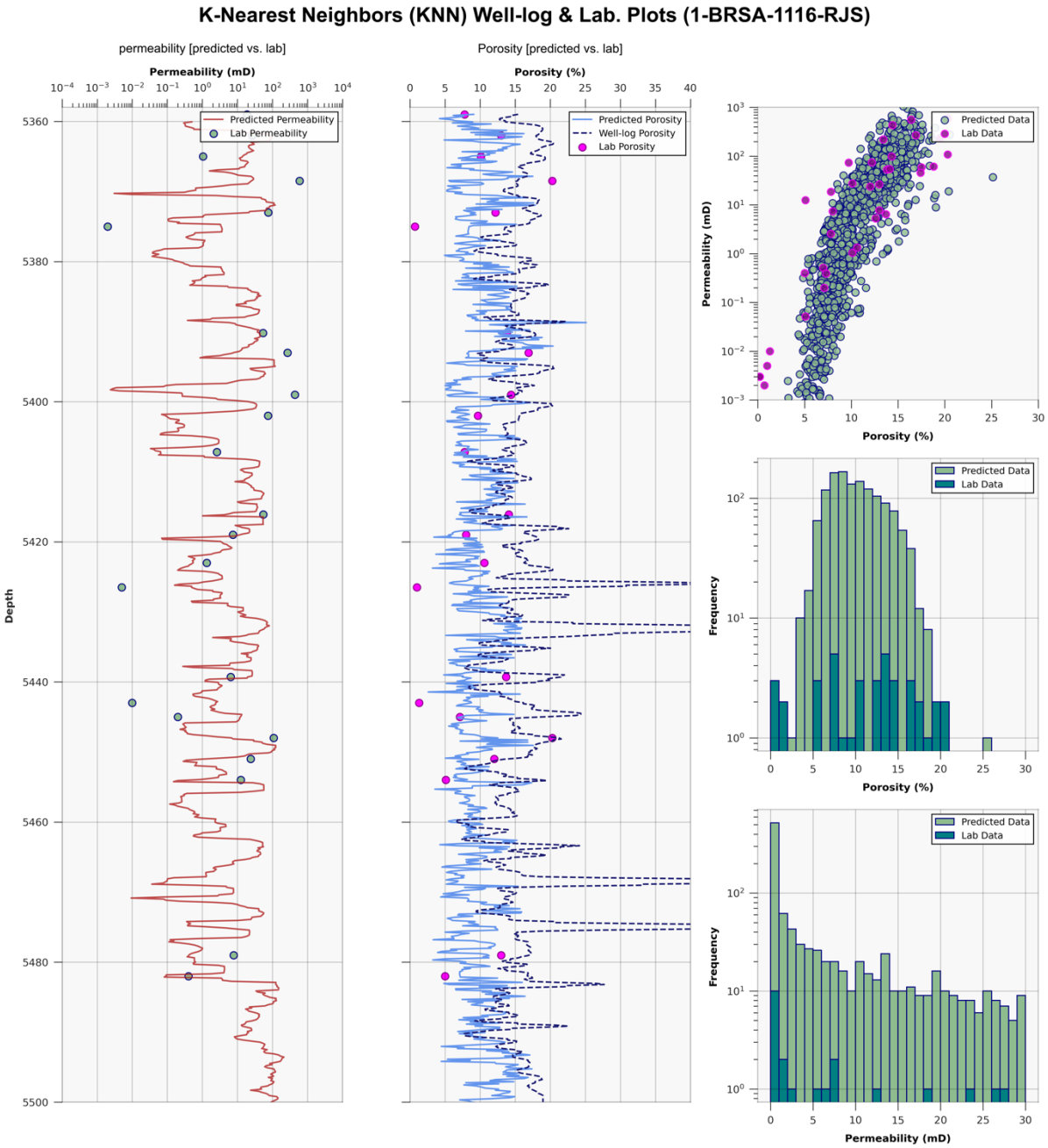

Figure 4-35: Comparing the match between the predicted & laboratory measurement of permeability and porosity, Comparing scatter plot permeability-porosity of predicted and laboratory data, Histogram of Porosity and Permeability distribution of predicted and laboratory data, for the GB Model, applied to a different well (3-BRSA-1116-RJS) not included in training set (for Permeability Prediction Timur-Coates Correlation has been used): (From left to right) 1st track: Predicted and laboratory Permeability (mD). 2nd track: Predicted and Measured Porosity (%). 3rd track: Scatter plot of permeability-porosity of predicted and laboratory data (mD vs. %), 4th track: Histogram of Porosity of predicted and laboratory data (%), 5th track Histogram of Permeability of predicted and laboratory data (mD)

The results of two other wells, 3-BRSA-944A-RJS and 9_ITP_RJS is shown in the Appendix 02.

4.2 Part two

4.2.1 Evaluation of different model performances using the test dataset

Machine learning methods have emerged as powerful tools for estimating permeability directly from well log data. These methods leverage the wealth of information contained in well logs to predict permeability, thereby bypassing the need for costly and time-consuming laboratory measurements. Machine learning models can be trained on a dataset comprising well log data as inputs and laboratory-measured permeability as outputs. Once trained, these models can predict permeability for new, unseen well log data, providing a rapid and cost-effective means of permeability estimation. This approach holds great promise for enhancing our understanding of reservoir properties and optimizing oilfield development strategies.

4.2.1.1 Scatter Plot of Predicted and Measured permeability

Generation of a predictive model using machine learning for the direct prediction of permeability from well log data as predictors and permeability as target data; bypassing the NMR log to calculate permeability and predicting it directly. By employing the LSSVR (Least Squares Support Vector Regressor), we develop a model that correlates well log data to permeability. The below scatter plot demonstrates how predicted and measured permeabilities correlate with each other. The scatter plots of predicted and measured permeability can be found in figures 4-36.

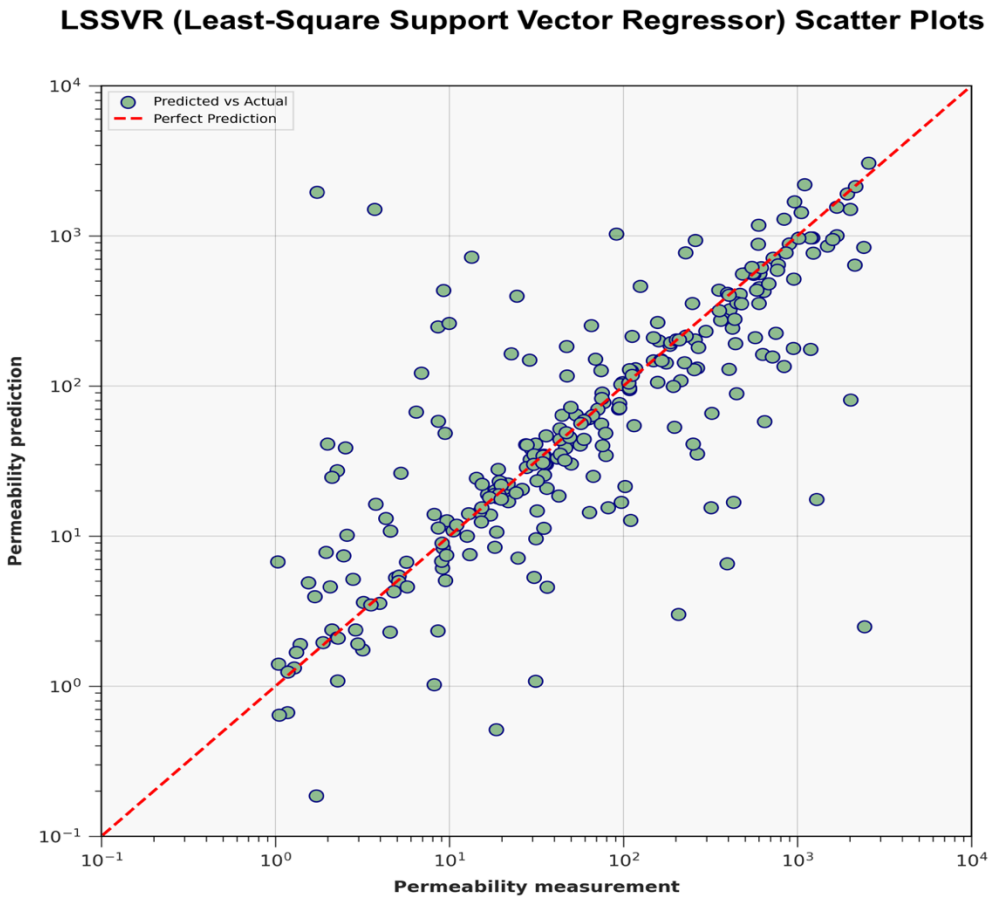

Figure 4-36: Scatter plots of predicted versus measured Permeability (mD), for the LSSVR Model, applied to four wells [(3-BRSA-944A-RJS), (1-BRSA-1116-RJS), (9-ITP-RJS), (3-BRSA-1215-RJS)] included in training phase.

4.2.1.2 Histogram of Predicted and Measured permeability

In this section, the histograms of LSSVR for both Laboratory Permeability and Predicted Permeability depict the frequency distribution of permeability values within specific ranges. This graphical representation is crucial for assessing the model's performance, as it visually illustrates how well the predicted permeability values align with the actual laboratory measurements. The significance lies in the ability to observe whether the predicted values fall within the same or, at the very least, a comparable range, indicating acceptable results. Essentially, the histogram serves as a valuable tool for gauging the accuracy and reliability of the predictive model in capturing the inherent patterns and variations in permeability across the dataset. The histogram of predicted and measured permeability can be found in figures 4-37.

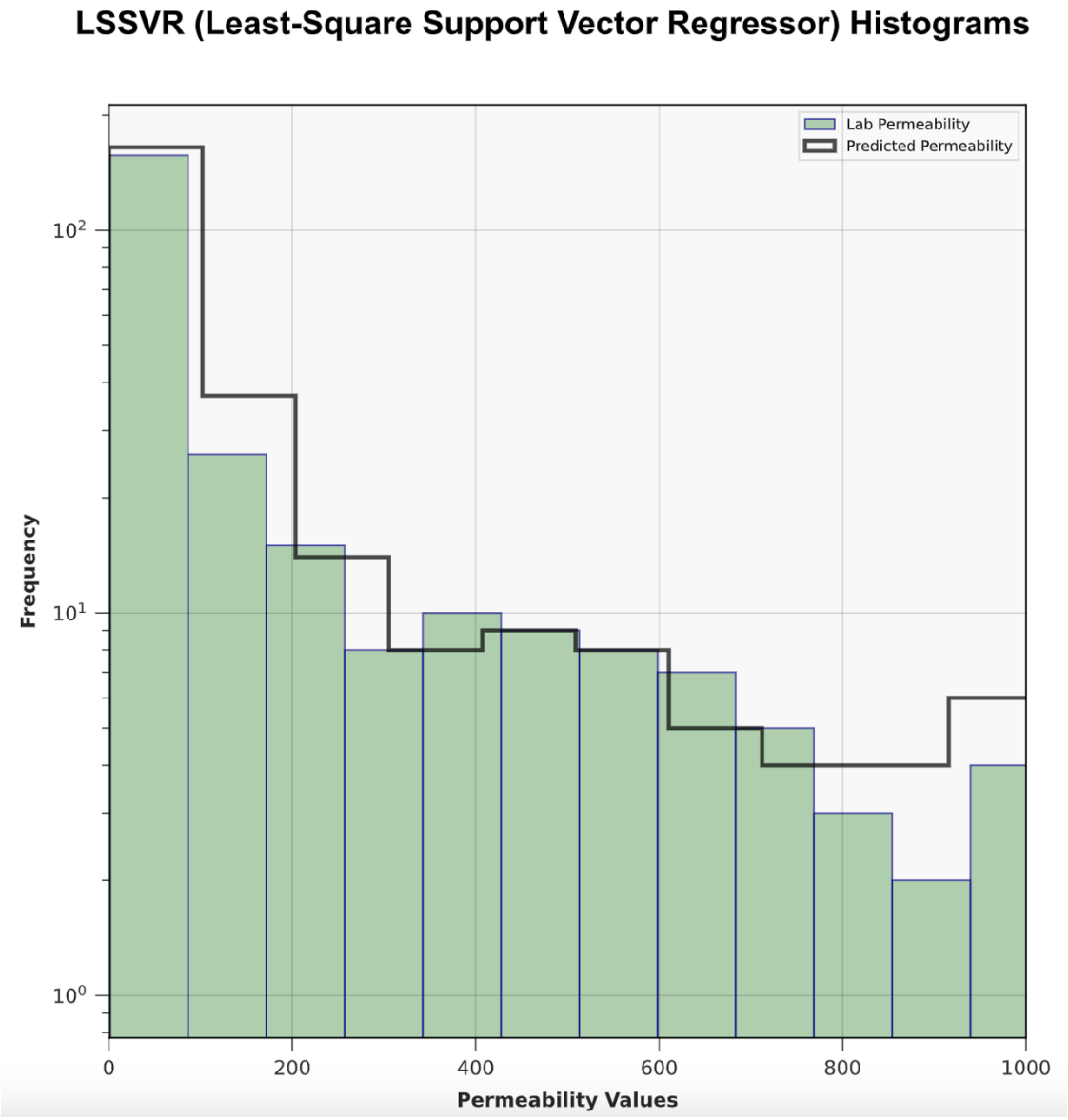

Figure 4-37: Comparative Histogram Analysis: LSSVR Predicted Permeability vs. Lab Permeability. Assessing the Frequency Distribution to Validate Model Performance

4.3 Using different regression metrics for model Evaluation

To assess the effectiveness of the RF, GB, and K-NN models utilized in our thesis, we employ several regression metrics. These metrics include Root Mean Squared Error (RMSE), Mean Squared Error (MSE), and Mean Absolute Error (MAE), the Coefficient of Determination (R2). Through the analysis of these metrics, we can evaluate and compare the performance of the models, allowing us to gauge their accuracy and predictive capabilities.

The Root Mean Squared Error (RMSE) is calculated by taking the square root of the Mean Squared Error (MSE). This metric measures the average squared difference between the predicted and measured values. The RMSE ranges between 0.00 and positive infinity, with a smaller value indicating a better fit. It is important to note that the RMSE uses the same scale as the measured data, allowing for easier interpretation and comparison.

$$RMSE = \sqrt{\frac{\sum_i^N (y_{Predicted} - y_{Measured})^2}{N}}$$

Equation 4-1

Similarly, the MSE represents the average squared difference between predicted and measured values, but it is not square rooted. Like the RMSE, the MSE ranges from 0.00 to positive infinity, with a smaller value indicating a better fit. However, the MSE is expressed in the squared units of the measured data, emphasizing the squared difference between the predictions and the actual values.

$$MSE = \frac{\sum_i^N (y_{Predicted} - y_{Measured})^2}{N}$$

Equation 4-2

The Mean Absolute Error (MAE) calculates the average absolute difference between the predicted and measured values. It is a metric that ranges from 0.00 to positive infinity, with a lower value indicating a better fit. The MAE uses the same scale as the measured data, making it easier to interpret and compare. Unlike the squared differences in MSE and RMSE, the MAE considers the absolute differences, giving equal weight to both positive and negative deviations from the actual values. Therefore, the MAE provides a measure of the average magnitude of errors in the predictions, without considering their direction.

$$MAE = \frac{\sum_i^N |y_{Predicted} - y_{Measured}|}{N}$$

Equation 4-3

The Coefficient of Determination (R2) is a numerical measure that indicates the predictive strength of a model, ranging between negative infinity and 1.00. A higher value closer to 1.00 signifies a stronger prediction. Conversely, a negative R2 value suggests that the model does not align with the data trend, which can occur in non-linear regression models. R2 is typically expressed as a percentage and is

calculated using a specific formula to quantify the proportion of the variance in the dependent variable that can be explained by the independent variables.

$$R^2 = 1 - \frac{\sum_i^N (y_{Predicted} - y_{Measured})^2}{\sum_i^N (y_{Predicted} - y_{Mean})^2}$$

Equation 4-4

4.3.1 Regression metrics evaluation on the test data

The succinct presentation of results encompasses the R2, MSE, RMSE, and MAE scores, obtained through the utilization of the most effective hyperparameter configuration. These scores are explicitly associated with predictions made on the test dataset, with a focus on the Random Forest (RF), Gradient Boosting (GB), and k-Nearest Neighbors (KNN) models. The R2, MSE, RMSE, and MAE scores of different porosities for different models of the test dataset can be found in table 4-1 and figures 4-38, 4-39, and 4-40.

Regression Metrics of RF, GB, and K-NN Models for estimation of Porosity on the Test Dataset

Machine Learning Algorithm	Types of Predicted Porosity	R^2 (%)	RMSE (m^3/m^3)	MSE $(m^3/m^3)^2$	MAE (m^3/m^3)
Random Forest (**RF**)	Effective	84.77	0.0179	0.0003	0.0117
	Free Fluid	86.39	0.016	0.0003	0.0105
	Total	80.19	0.018	0.0003	0.012
Gradient Boosting (**GB**)	Effective	84.62	0.018	0.0003	0.0117
	Free Fluid	86.26	0.0161	0.0003	0.0105
	Total	78.56	0.0188	0.0004	0.0123
K-Nearest Neighbour (**K-NN**)	Effective	64.94	0.0271	0.0007	0.018
	Free Fluid	65.63	0.0254	0.0006	0.0164
	Total	56.01	0.0269	0.0007	0.0181

Table 4-1: Random Forest (RF), Gradient Boosting (GB), and K-Nearest Neighbour Regression (K-NN) Metrics on Test Dataset: Comparative Analysis of Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Squared Error (MSE), and Coefficient of Determination (R^2) for Different Porosity Types.

Moreover, we provide bar plots that visually represent the significance of each input feature in predicting the output parameters. These graphical representations assist in discerning the most influential features contributing to the model's predictions.

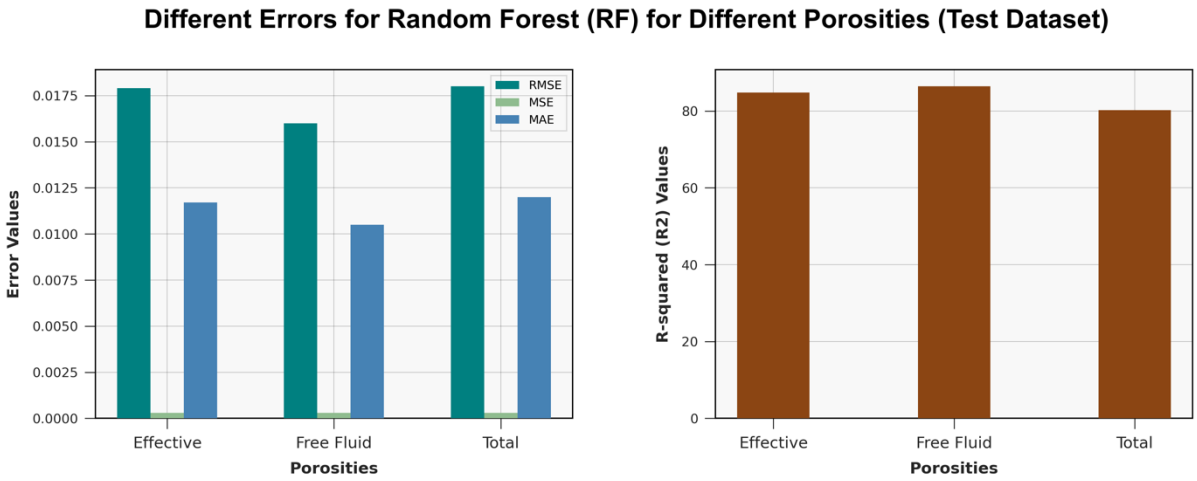

Figure 4-38: Random Forest Regression Metrics Histogram on Test Dataset: Comparative Analysis of Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Squared Error (MSE), and Coefficient of Determination (R^2) for Different Porosity Types.

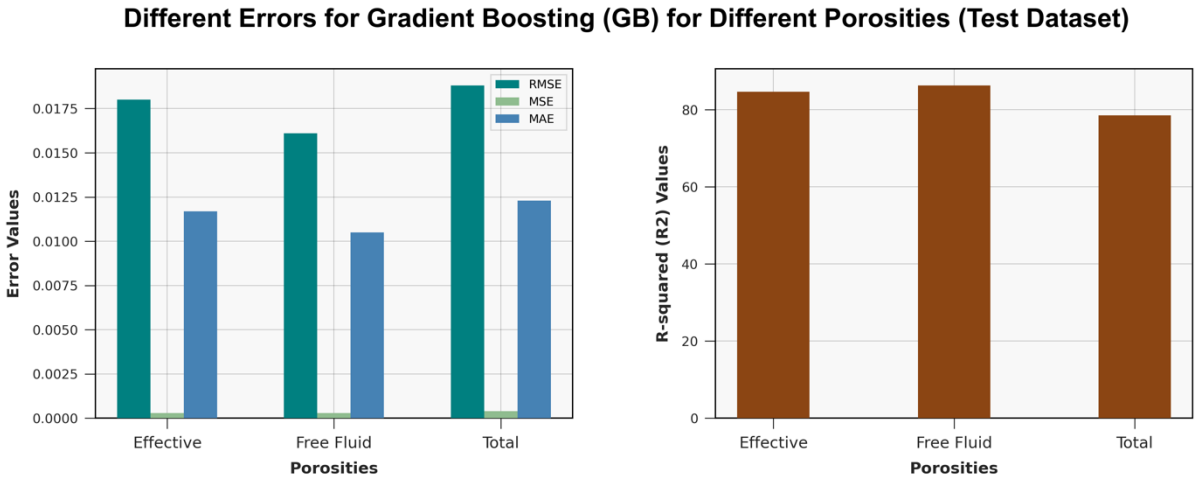

Figure 4-39: Gradient Boosting Regression Metrics Histogram on Test Dataset: Comparative Analysis of Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Squared Error (MSE), and Coefficient of Determination (R^2) for Different Porosity Types.

While both Random Forest (RF) and Gradient Boosting (GB) exhibit closely comparable results, the Random Forest model marginally outperforms the Gradient Boosting model, displaying superior outcomes with a higher coefficient of determination and lower root mean square error (RMSE). In contrast, the K-Nearest Neighbour (K-NN) model demonstrates the least accuracy, as indicated by the R^2 (Coefficient of Determination).

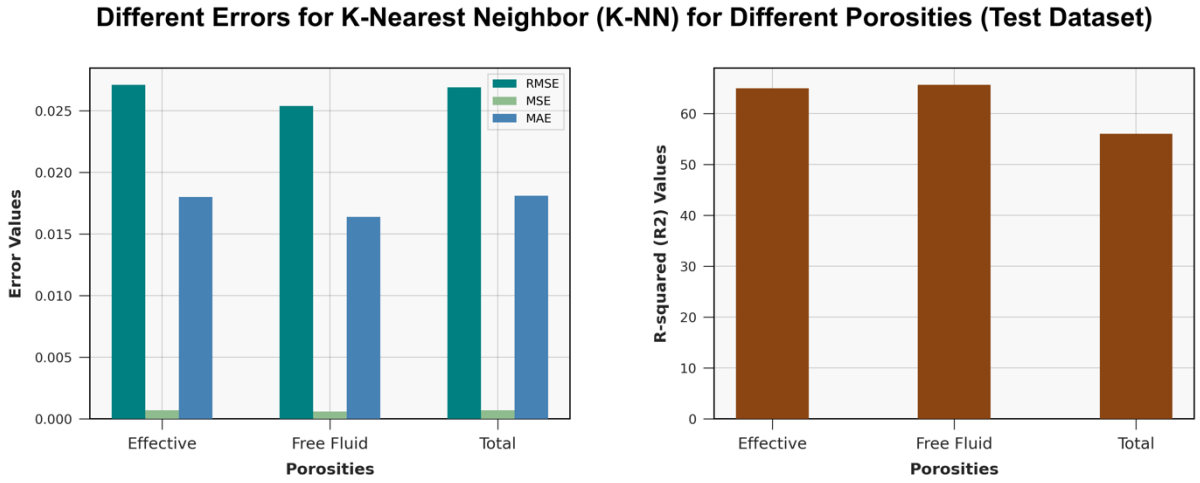

Figure 4-40: K-Nearest Neighbour Regression Metrics Histogram on Test Dataset: Comparative Analysis of Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Squared Error (MSE), and Coefficient of Determination (R^2) for Different Porosity Types.

The observed disparities in model performance may be attributed to the varying degrees of correlation between the well log data and NMR Free Fluid. Notably, all three models display competence in predicting Free Fluid Porosity, but the Random Forest model stands out with its slightly better predictive accuracy, emphasizing its efficacy in capturing the nuanced relationships within the dataset.

4.3.2 Regression metrics evaluation on the entire dataset

The concise summary of outcomes encapsulates the R2, MSE, RMSE, and MAE metrics achieved by employing the most optimal hyperparameter settings. These scores are specifically linked to predictions made across the entire dataset, emphasizing the performance of the Random Forest (RF), Gradient Boosting (GB), and k-Nearest Neighbors (KNN) models. The R2, MSE, RMSE, and MAE scores of different porosities for different models of the entire dataset can be found in table 4-2 and figures 4-41, 4-42, and 4-43.

Regression Metrics of RF, GB, and K-NN Models for estimation of Porosity on the Entire Dataset

Machine Learning Algorithm	Types of Predicted Porosity	R^2 (%)	RMSE (m^3/m^3)	MSE $(m^3/m^3)^2$	MAE (m^3/m^3)
Random Forest (**RF**)	Effective	95.31	0.0103	0.0001	0.006
	Free Fluid	95.69	0.0093	0.0001	0.0054
	Total	94.08	0.0103	0.0001	0.0061
Gradient Boosting (**GB**)	Effective	96.85	0.0084	0.0001	0.004
	Free Fluid	97.14	0.0161	0.0001	0.0036
	Total	95.62	0.0089	0.0001	0.0044
K-Nearest Neighbour (**K-NN**)	Effective	93.5	0.0121	0.0001	0.0036
	Free Fluid	93.52	0.0114	0.0001	0.0033
	Total	91.97	0.012	0.0001	0.0036

Table 4-2: Random Forest (RF), Gradient Boosting (GB), and K-Nearest Neighbour Regression (K-NN) Metrics on Entire Dataset: Comparative Analysis of Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Squared Error (MSE), and Coefficient of Determination (R^2) for Different Porosity Types.

Furthermore, the presentation includes graphical representations in the form of bar plots, effectively depicting the significance of each input feature in predicting the output parameters. These visual aids serve the purpose of discerning the pivotal features that play a substantial role in influencing the model's predictions.

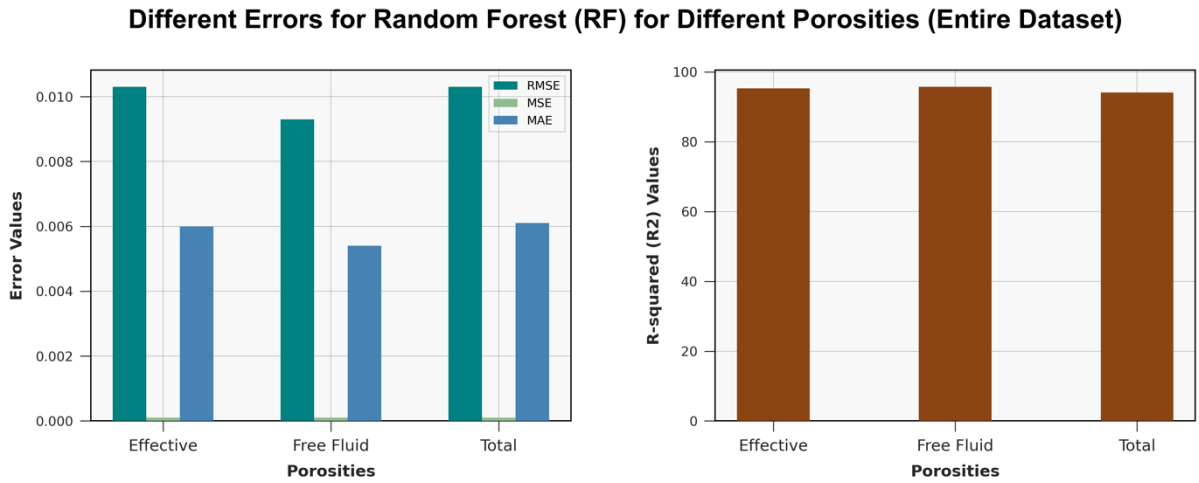

Figure 4-41: Random Forest Regression Metrics Histogram on Entire Dataset: Comparative Analysis of Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Squared Error (MSE), and Coefficient of Determination (R^2) for Different Porosity Types.

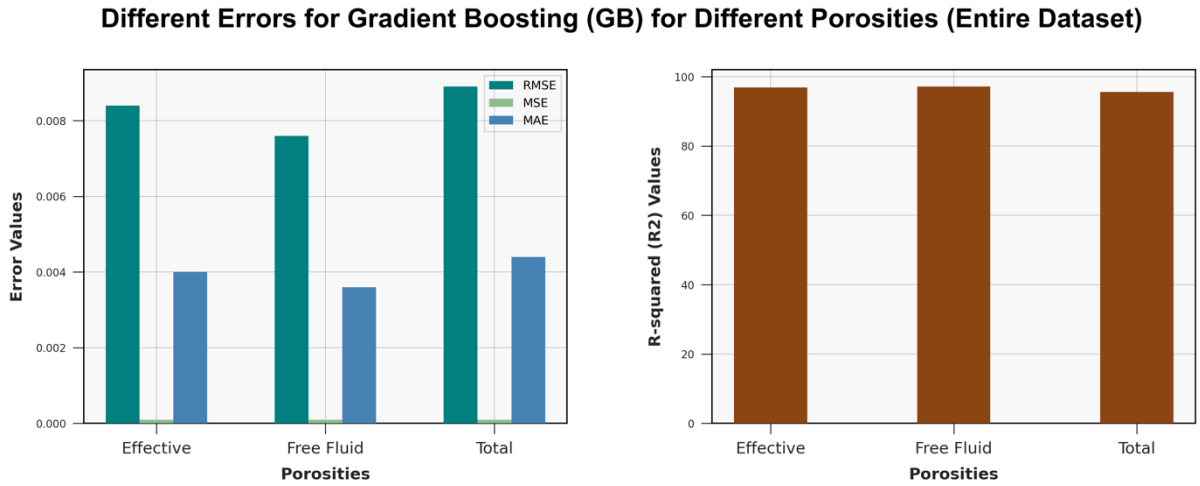

Figure 4-42: Gradient Boosting Regression Metrics Histogram on Entire Dataset: Comparative Analysis of Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Squared Error (MSE), and Coefficient of Determination (R^2) for Different Porosity Types.

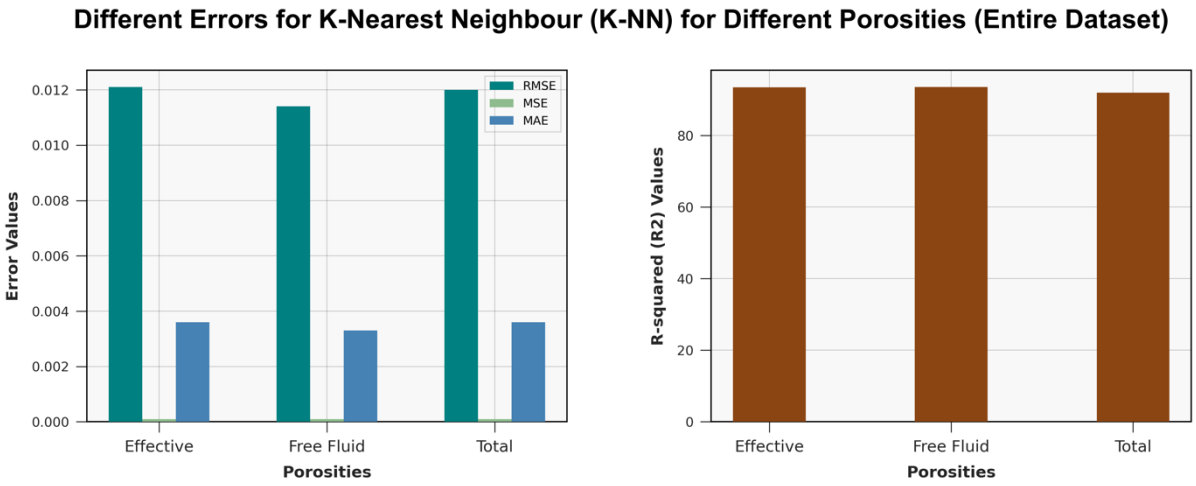

Figure 4-43: K-Nearest Neighbour Regression Metrics Histogram on Entire Dataset: Comparative Analysis of Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Squared Error (MSE), and Coefficient of Determination (R^2) for Different Porosity Types.

Compared to the test dataset, the entire dataset yields significantly improved results across all three models, showcasing a notable enhancement in the coefficient of determination by approximately 30-40%. Among the models, Gradient Boosting stands out as the top performer, surpassing the other two. Consistent with the test dataset, the models continue to demonstrate a superior performance in predicting free fluid porosities compared to effective total porosities.

This robust performance on the entire dataset underscores the generalizability and effectiveness of the models in capturing the underlying patterns within the broader dataset. The heightened accuracy, particularly in the case of Gradient Boosting, signifies its prowess in handling a more extensive range of data, emphasizing its potential for reliable predictions in real-world scenarios. Furthermore, the continued trend of better performance in predicting free fluid porosities suggests a consistent strength in capturing the nuances of this specific porosity type across various datasets.

4.3.3 Regression metrics evaluation on the different datasets (1-BRSA-1116-RJS)

To assess the models'generalizability, a separate well data set that was not used in training was employed. The models were evaluated using R2, MSE, RMSE, and MAE metrics obtained with the optimal hyperparameter settings. This evaluation method provided a realistic assessment of the models'generalizability. The R2, MSE, RMSE, and MAE scores of different porosities for different models of the different dataset (not included in training the models) can be found in table 4-3 and figures 4-44, 4-45, and 4-46.

Regression Metrics of RF, GB, and K-NN Models for estimation of Porosity on the Different Dataset

Machine Learning Algorithm	Types of Predicted Porosity	R^2 (%)	RMSE (m^3/m^3)	MSE $(m^3/m^3)^2$	MAE (m^3/m^3)
Random Forest (**RF**)	Effective	-52.02	0.0695	0.0048	0.0504
	Free Fluid	-68.78	0.045	0.002	0.036
	Total	-34.84	0.0672	0.0045	0.0473
Gradient Boosting (**GB**)	Effective	-58.65	0.071	0.005	0.0512
	Free Fluid	-84.86	0.0471	0.0022	0.0386
	Total	-43.79	0.0694	0.0048	0.0493
K-Nearest Neighbour (**K-NN**)	Effective	-144.96	0.0882	0.0078	0.0663
	Free Fluid	-202.86	0.0603	0.0036	0.0491
	Total	-106.27	0.0831	0.0069	0.061

Table 4-3: Random Forest (RF), Gradient Boosting (GB), and K-Nearest Neighbour Regression (K-NN) Metrics on Different Dataset (Not Included in training phase): Comparative Analysis of Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Squared Error (MSE), and Coefficient of Determination (R^2) for Different Porosity Types.

the presentation includes graphical representations in the form of bar plots are shown below:

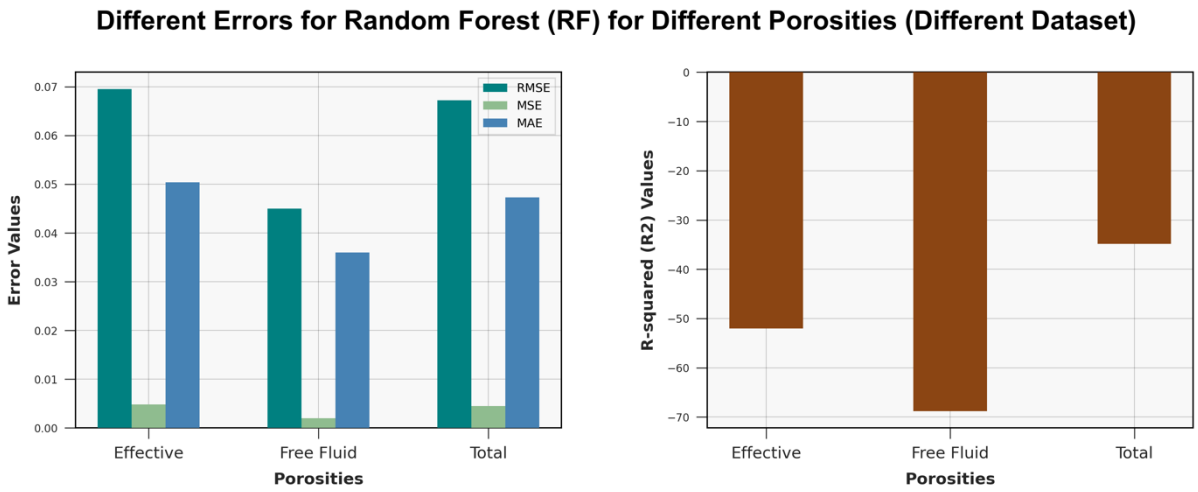

Figure 4-44: Random Forest Regression Metrics Histogram on Different Dataset (Not Included in training phase): Comparative Analysis of Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Squared Error (MSE), and Coefficient of Determination (R^2) for Different Porosity Types.

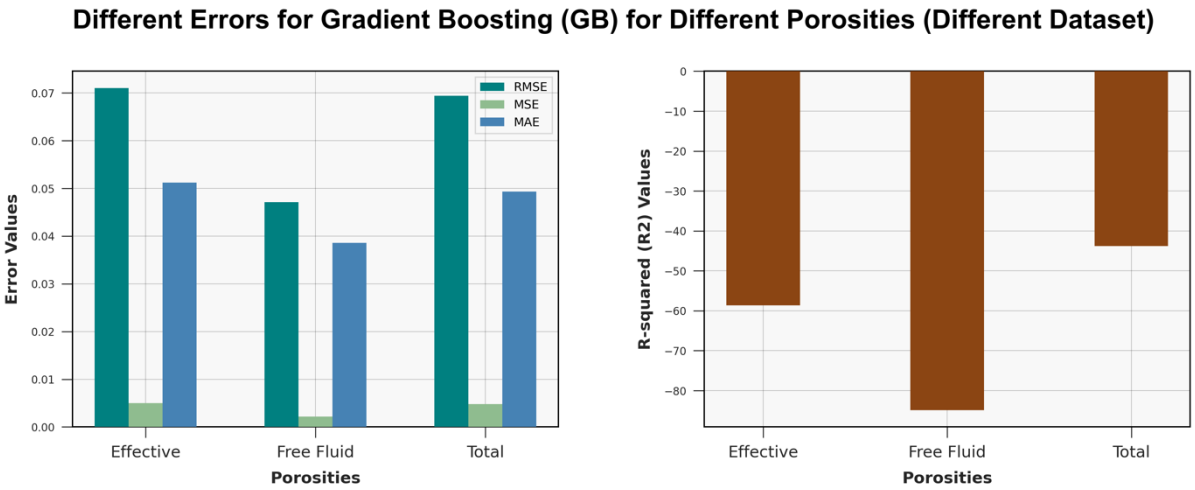

Figure 4-45: Gradient Boosting Regression Metrics Histogram on Different Dataset (Not Included in training phase): Comparative Analysis of Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Squared Error (MSE), and Coefficient of Determination (R^2) for Different Porosity Types.

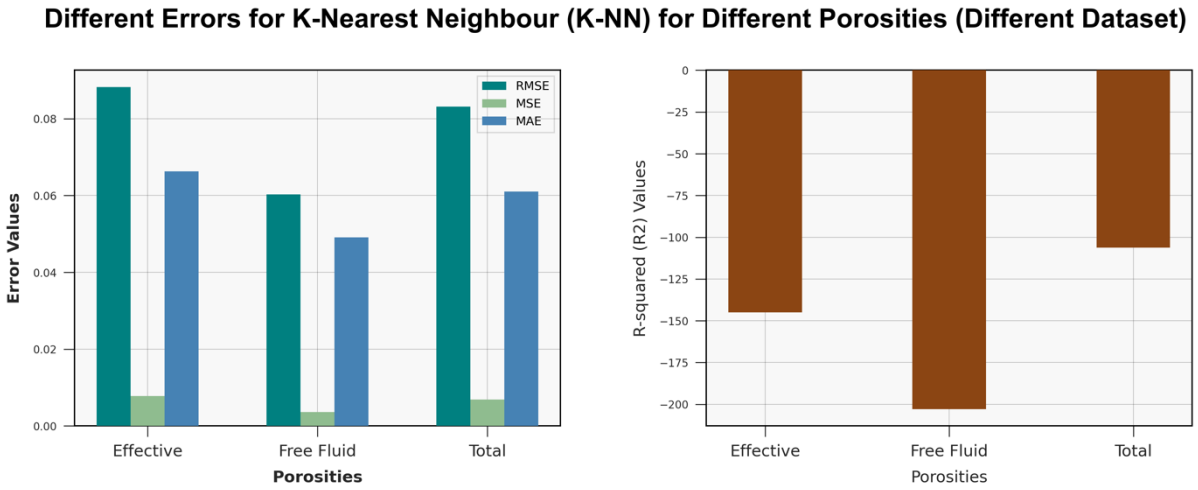

Figure 4-46: K-Nearest Neighbour Regression Metrics Histogram on Different Dataset (Not Included in training phase): Comparative Analysis of Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Squared Error (MSE), and Coefficient of Determination (R^2) for Different Porosity Types.

Apply a different dataset not including in training the model, show negative coefficient of determination for all models and for different types of porosities. This indicates that the models struggle to accurately predict the output parameters on the test well. There are several reasons for this lack of accuracy in the validation phase. Firstly, the training dataset partly consists of non-reservoir rock in some wells, making it challenging for the models to generalize to the test dataset, which is almost entirely composed of reservoir rock (90%), even if they are made of the same lithology.

Secondly, the training dataset itself is limited. The small size of the training dataset restricts the model's ability to capture the underlying relationships between the input features and the output parameters, accurately.

5 Conclusion

In this study, the main objective was to develop machine learning models for the prediction of petrophysical data using Random Forest (RF), Gradient Boosting (GB), K-Nearest Neighbors (K-NN), and Least Square Support Vector Regressor (LSSVR) algorithms. The models were trained using data indirectly obtained from predicted NMR well-log data and laboratory permeability. The evaluation of the models was performed on different types of porosities (free fluid, effective, total) using both test and entire datasets. Additionally, the study compared the predicted permeability using predicted NMR and Timur-Coates correlation with laboratory permeability. Finally, LSSVR was used to predict permeability directly from well-log data.

It was concluded that:

1. In the first part of the thesis, three different models (Random Forest, Gradient Boosting, and K-Nearest Neighbour) were evaluated using test dataset, entire dataset, and a different dataset not included in the training model. It has been shown that there are acceptible and cohesive scatter plots (measured vs. predicted) and log plots (measured vs. predicted) for each models.
2. Generally, Random Forest and Gradient Boosting showed a better results in comparison to K-Nearest Neighbour model; The test dataset showed that the Random Forest model slightly outperformed the Gradient Boosting model in terms of coefficient of determination and root mean square error. The entire dataset depicted that all three models showed significantly improved results compared to the test dataset. However, application of different dataset, not included in the model training, all models yielded negative coefficients of determination for different types of porosities.
3. However, the K-Nearest Neighbour model demonstrated the least accuracy. The reseaon behind this is the limititaion of K-Nearest Neighbour in dealing with high number of data (the number of data in the first part is around 10000).
4. Application of Random Forest, Gradient Boosting, and K-Nearest Neighbour on each well (expecially for well: 3-BRSA-1215-RJS), separately, show good matches for predicted and measured permeability anfd porosities using Timur-Coates correlations. Other wells (3-BRSA-944A-RJS and 9-lTP-RJS) also depicted close matches and acceptible results, while application of the differnet well not included in the developement of model (1-BRSA-1116-RJS) show less accurate results, the results are totally acceptible.
5. In the second part of the thesis, an acceptible close matches has been shown; direct permeability prediction from basic well-log data is an valuable predciton and a big step in reservoir characterization. However, there are some inaccuracies in the results mainly due to the lack of sufficient data.

Below are some notes pertaining to existing inaccuracies in the thesis, as well as suggestions for improving the research and recommendations for potential future research:

- There are some inaccuracies in the results, especially in the second part that predicts permeability directly from basic well-log data. This is mainly due to the lack of sufficient data; the training data comprises 191 samples, and the testing data consists of 82 samples, giving us a total of 273 samples. The higher the number of data, the better, more coherent, and accurate the machine

learning model. The reason for the low number of data is that the dataset for the second part comes from the concentration of well-log data and laboratory data, and in this case, laboratory data was the limiting factor due to a limited number of lab data.

- Depth shift in laboratory results show the better alignment of laboratory results and predicted data. This inaccuracy can be due to the differences between loggers and drillers depths are due to different stretch in the drilling string when drilling, and the wire line entered into the bore hole during wireline logging operations. This shift can suggest a more refined synchronization, indicating an enhanced accuracy and reliability in the predictive model.
- While the application of the Timur-Coates correlation has yielded good and generally acceptable results in predicting permeability, it is essential to note that the model is deterministic and may not be exact for all cases. Originally developed for sandstone reservoirs, its application to carbonate reservoirs could potentially introduce errors. One consideration to address this limitation involves modifying the model based on geological features, specifically by adjusting the coefficients of the Timur-Coates correlation (a, b, and c) to better suit carbonate reservoir conditions.
- Due to the inaccuracy of the Timur-Coates correlation that has been mentioned, and considering the fact that it serves as the link between the two results – predicted permeability (first part) and laboratory permeability – it could be critically asserted that in instances where the predicted permeability deviates from the laboratory permeability, the predicted results may not necessarily be inaccurate. Instead, it is plausible that the laboratory measurements themselves are inaccurate.

Appendix

Appendix 01: (Histogram & Distributions plots)

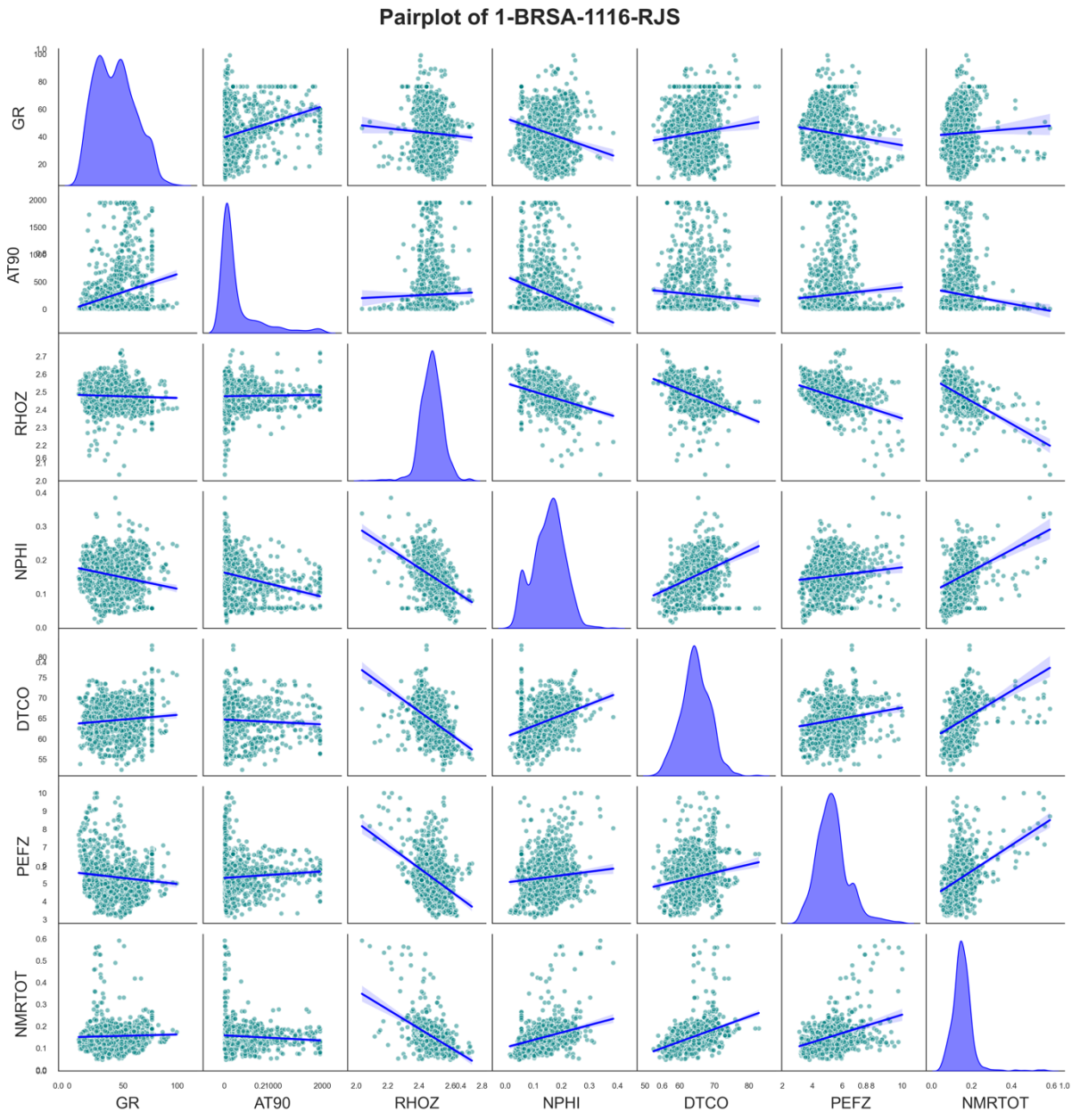

Figure A-1: Seaborn Pair plots of well: 1-BRSA-1116-RJS for different well log data (GR, AT90, RHOZ, NPHI, DTCO, PEFZ, and NMRTOT)

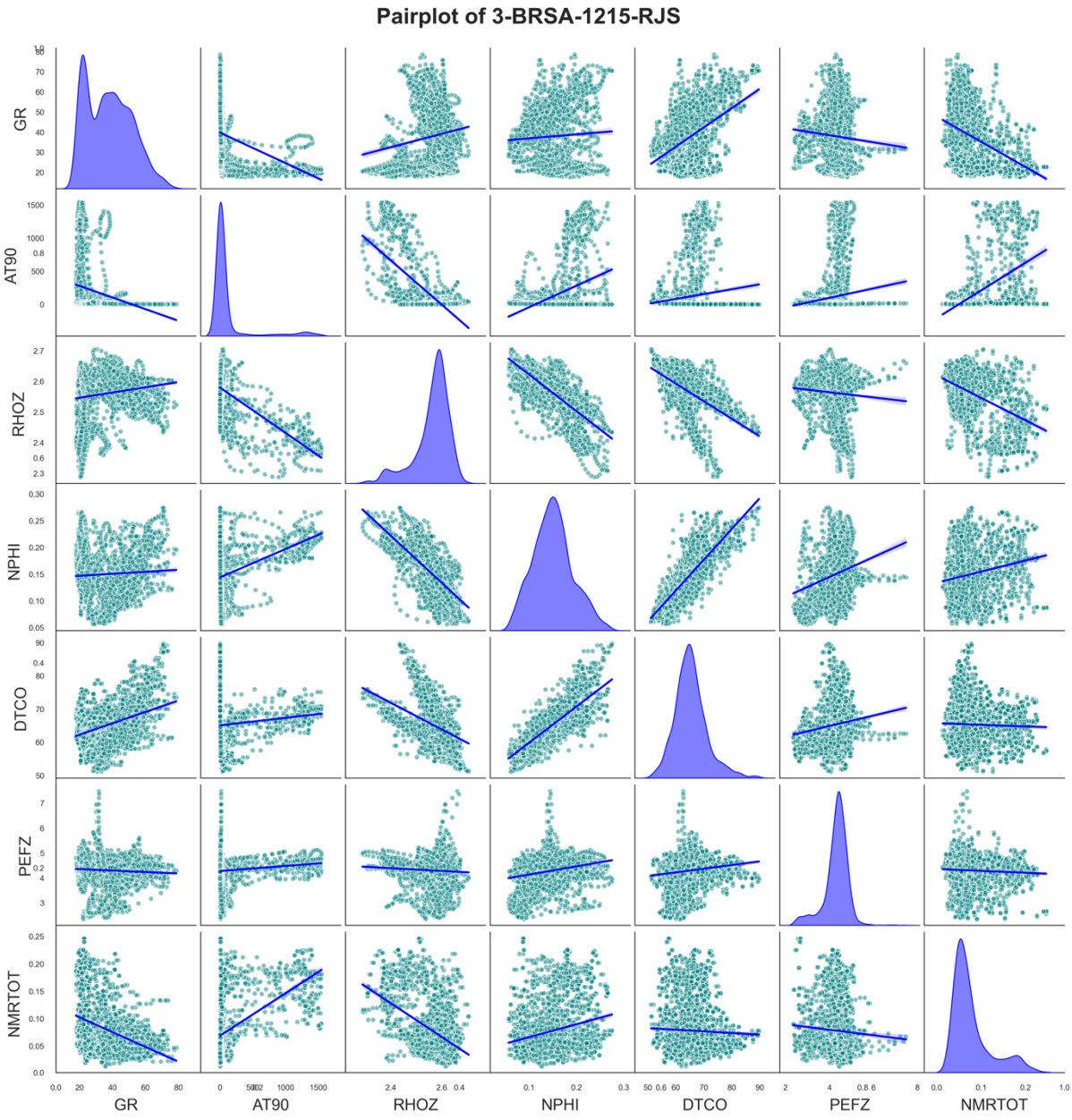

Figure A-2: Seaborn Pair plots of well: 3-BRSA-1215-RJS for different well log data (GR, AT90, RHOZ, NPHI, DTCO, PEFZ, and NMRTOT)

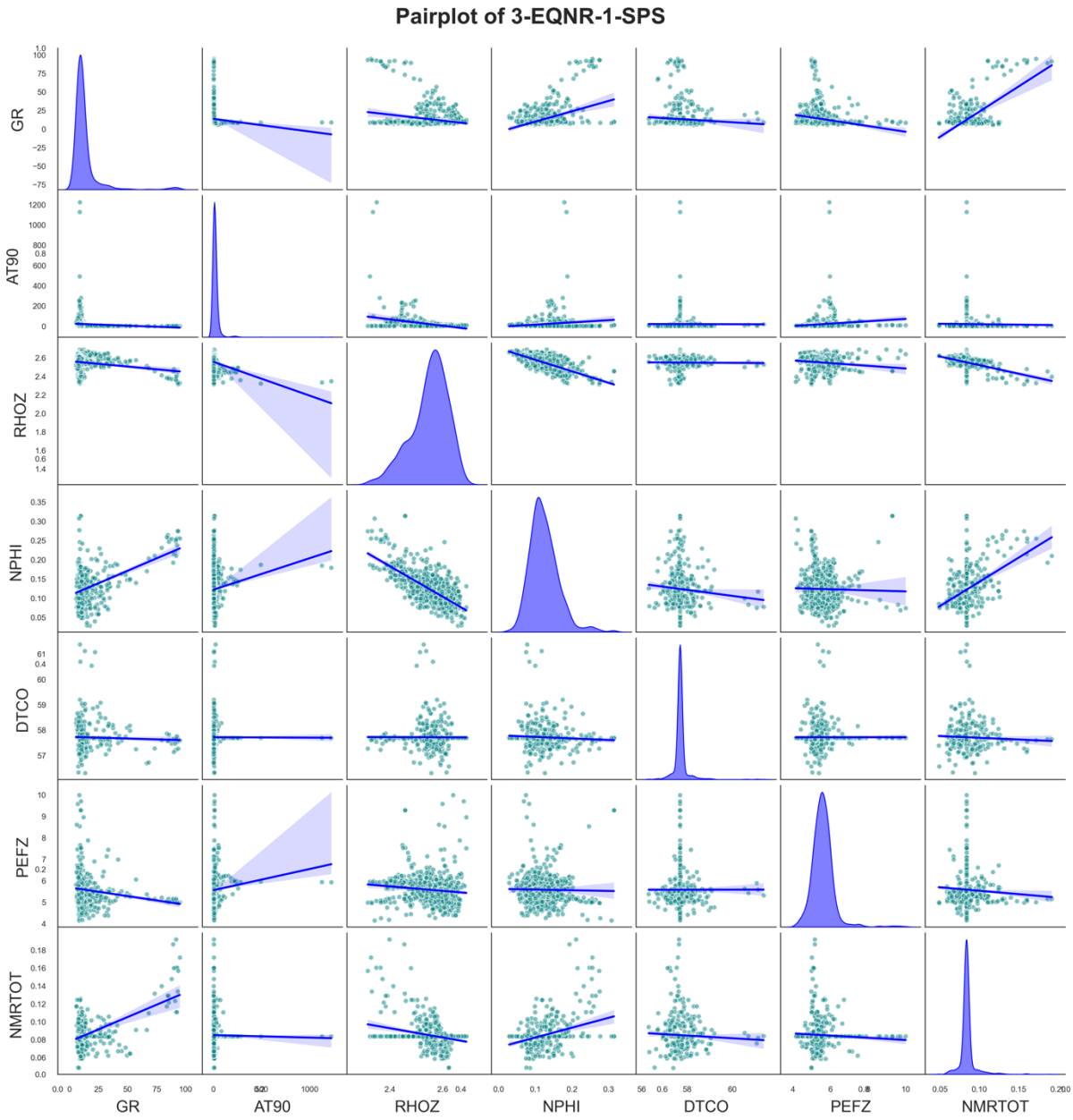

Figure A-3: Seaborn Pair plots of well: 3-EQNR-1-SPS for different well log data (GR, AT90, RHOZ, NPHI, DTCO, PEFZ, and NMRTOT)

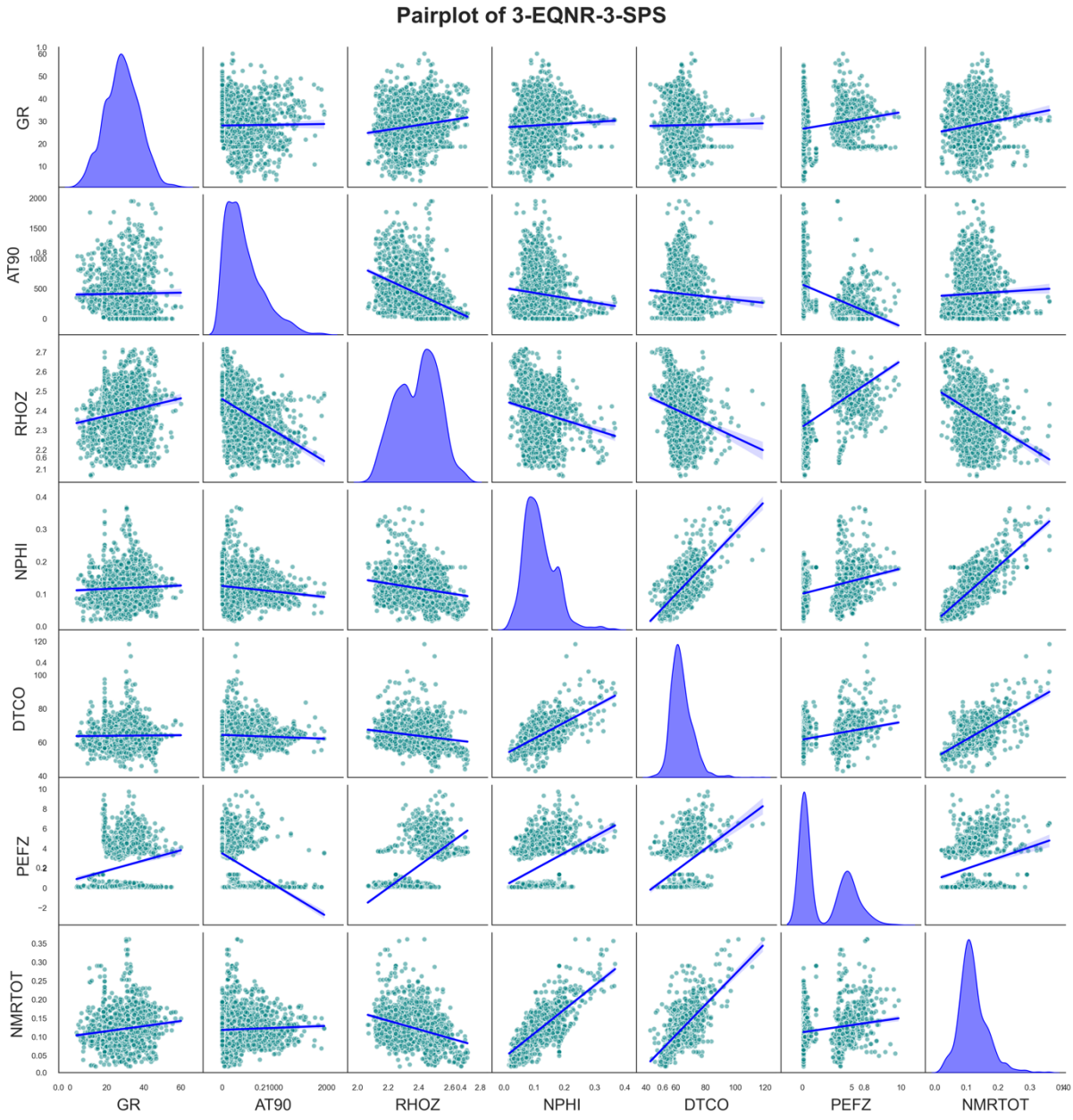

Figure A-4: Seaborn Pair plots of well: 3-EQNR-3-SPS for different well log data (GR, AT90, RHOZ, NPHI, DTCO, PEFZ, and NMRTOT)

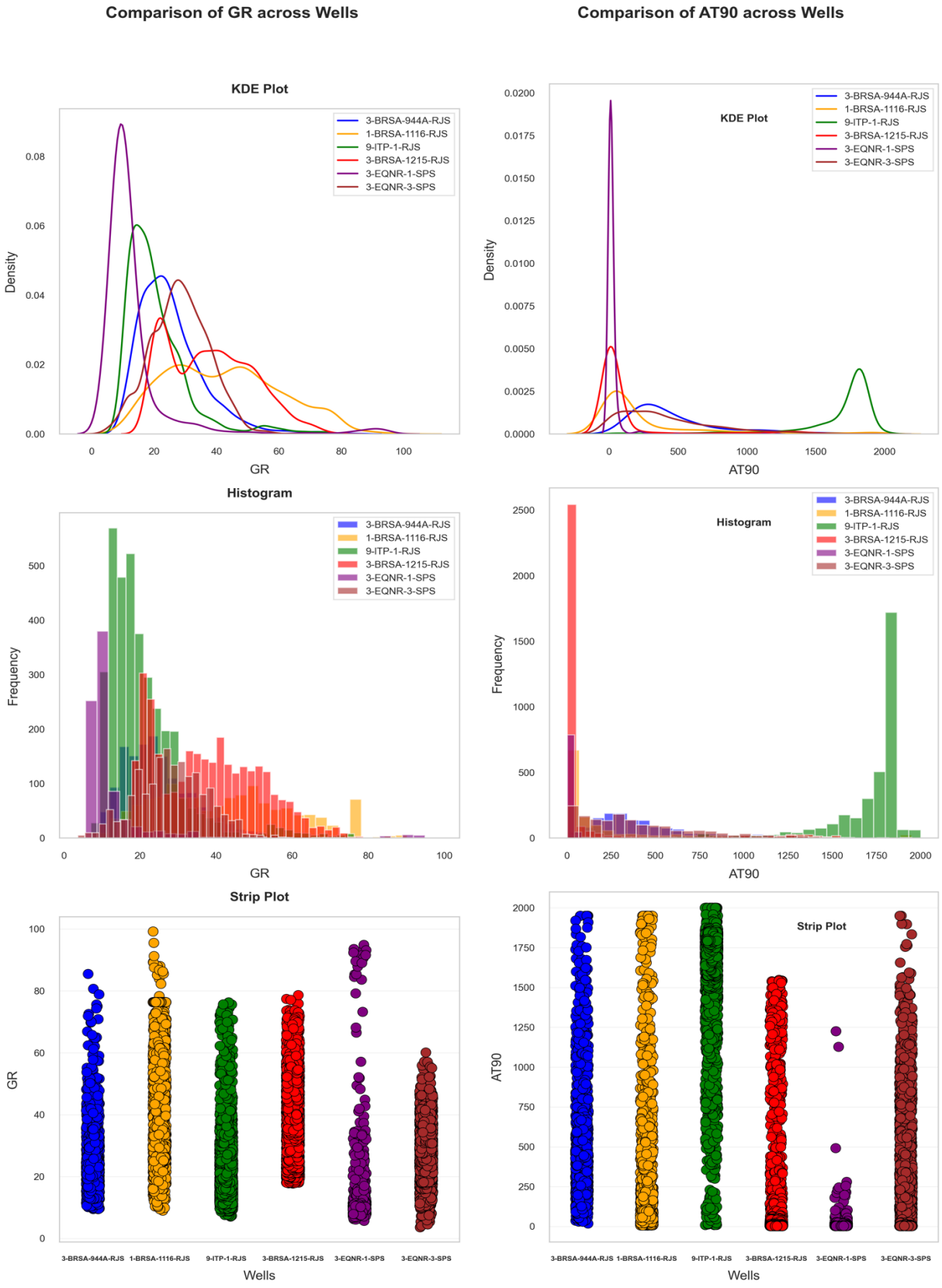

Figure A-5: Different Plots (KDE Plot, Histogram, and Strip) for GR on the left and AT90 on the right side. Each of these plots for different properties (GR & AT90) show the distribution and comparison the range of the parameters for six different wells.

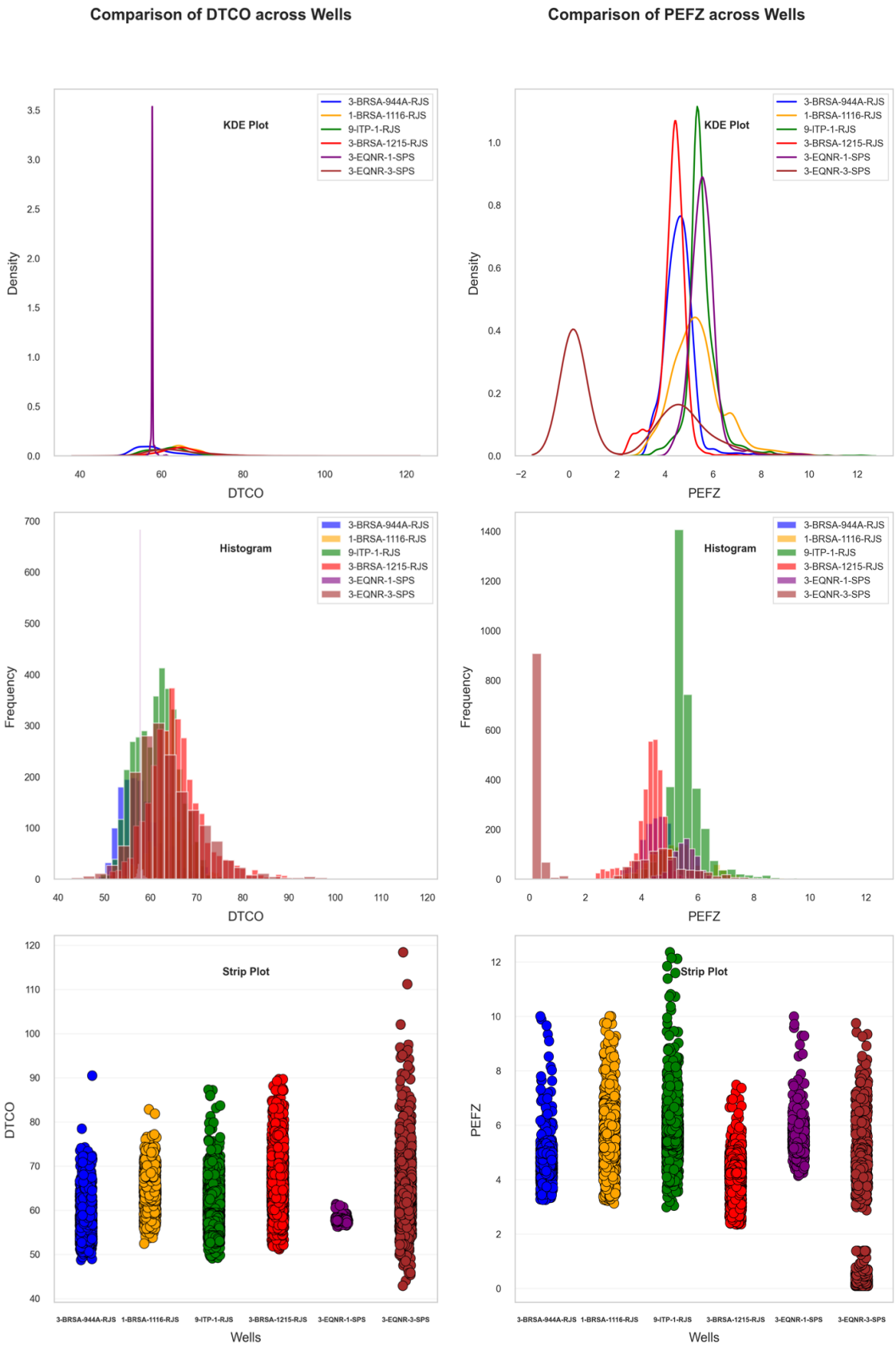

Figure A-6: Different Plots (KDE Plot, Histogram, and Strip) for DTCO on the left and PEFZ on the right side. Each of these plots for different properties (DTCO & PEFZ) show the distribution and comparison the range of the parameters for six different wells.

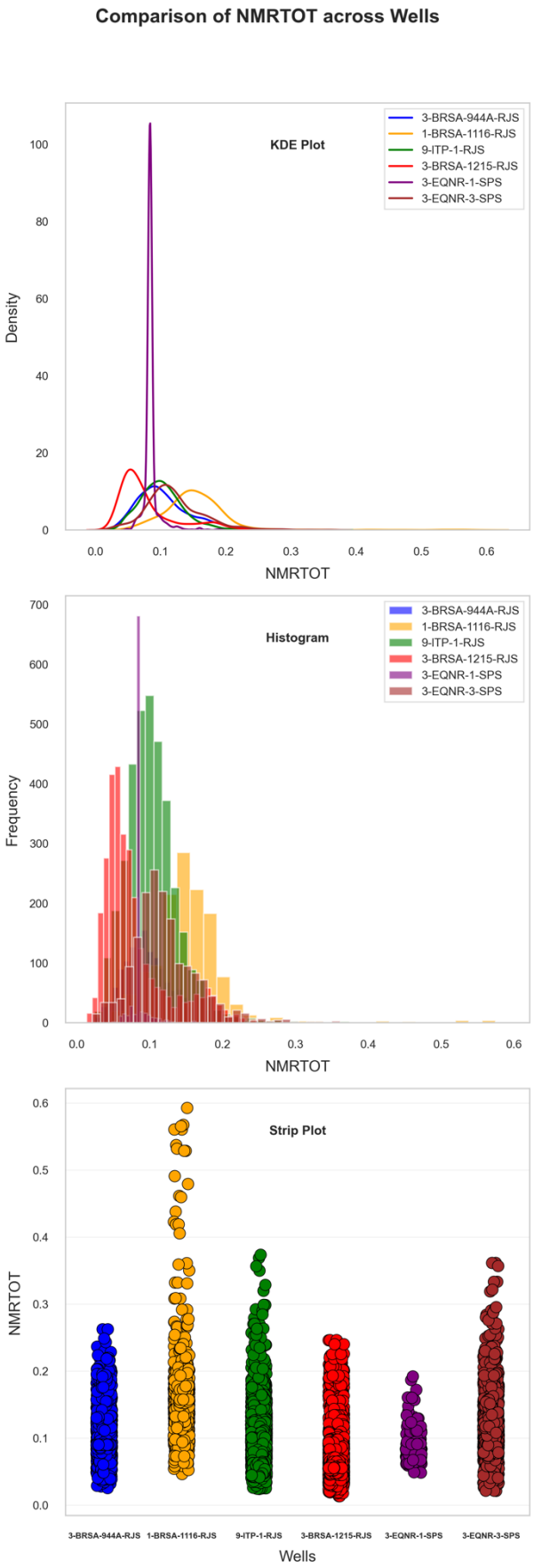

Figure A-7: Different Plots (KDE Plot, Histogram, and Strip) for NMRTOT show the distribution and comparison the range of the parameters for six different wells.

Appendix 02: (Well-log plots)

Evaluation of different model performances using the different dataset [3-BRSA-944A-RJS] (Not included in training-test)

Scatter plots

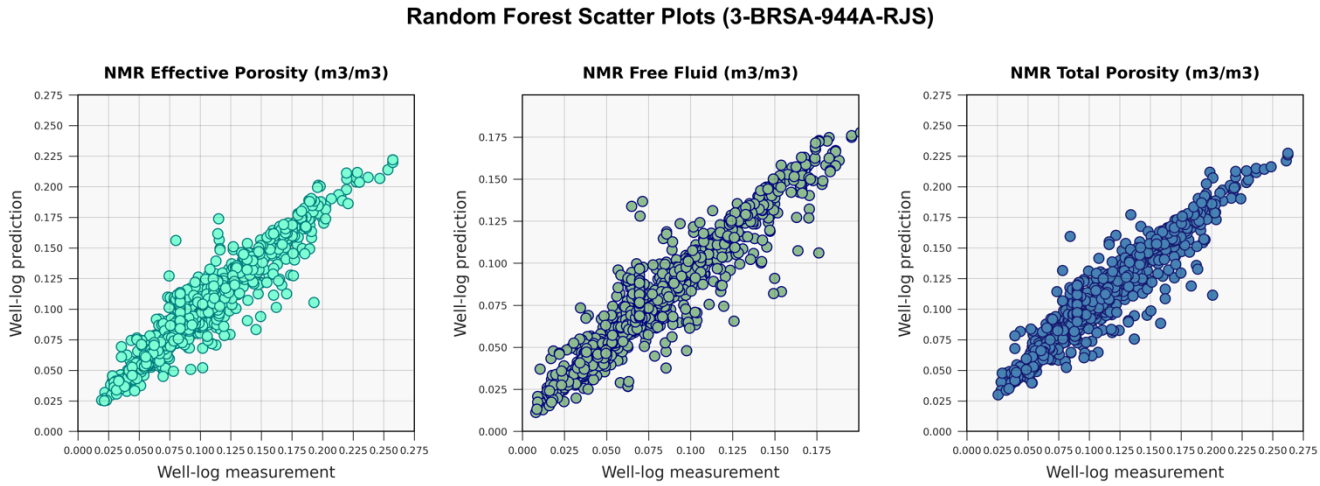

Figure A-8: Scatter plots of predicted versus measured NMR porosity, for the RF Model, applied to one of the well (3-BRSA-944A-RJS) included in training phase: Left plot: NMR Effective Porosity (m3/m3). Middle plot: NMR Free Fluid (m3/m3). Right plot: NMR Total Porosity (m3/m3).

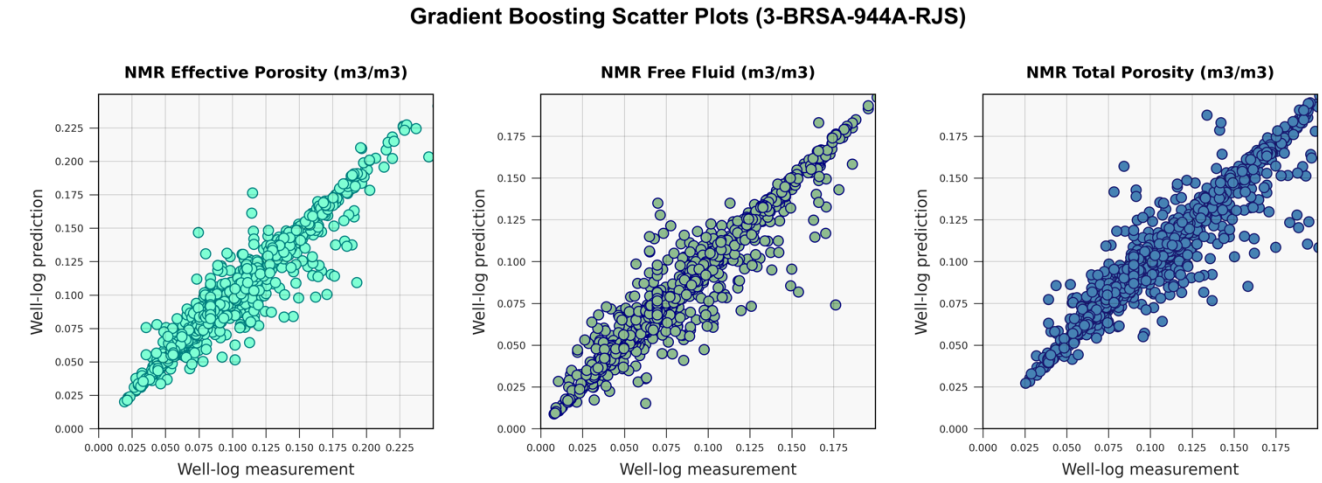

Figure A-9: Scatter plots of predicted versus measured NMR porosity, for the GB Model, applied to one of the well (3-BRSA-944A-RJS) included in training phase: Left plot: NMR Effective Porosity (m3/m3). Middle plot: NMR Free Fluid (m3/m3). Right plot: NMR Total Porosity (m3/m3).

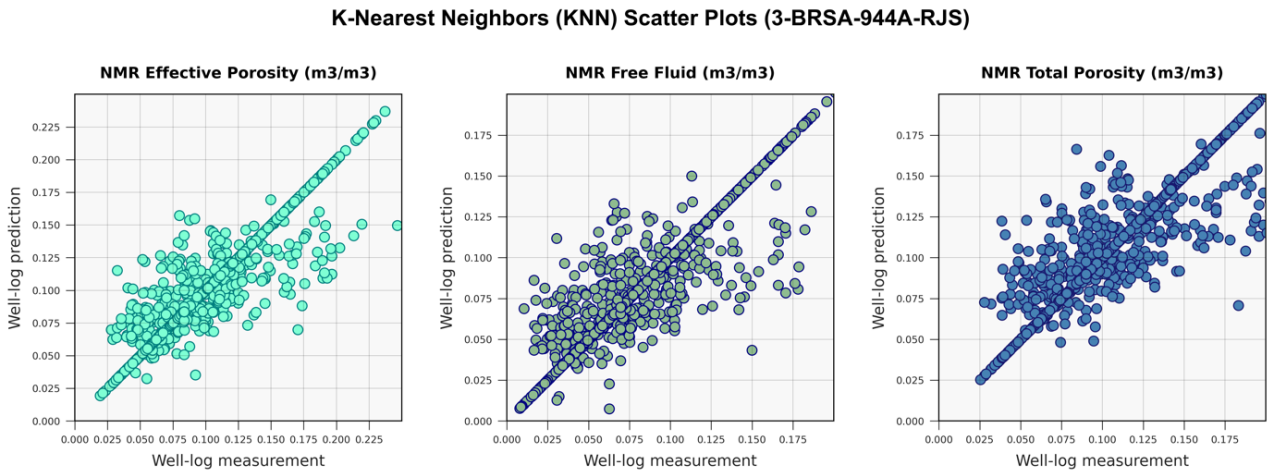

Figure A-10: Scatter plots of predicted versus measured NMR porosity, for the K-NN Model, applied to one of the well (3-BRSA-944A-RJS) included in training phase: Left plot: NMR Effective Porosity (m3/m3). Middle plot: NMR Free Fluid (m3/m3). Right plot: NMR Total Porosity (m3/m3).

Well-log Plots

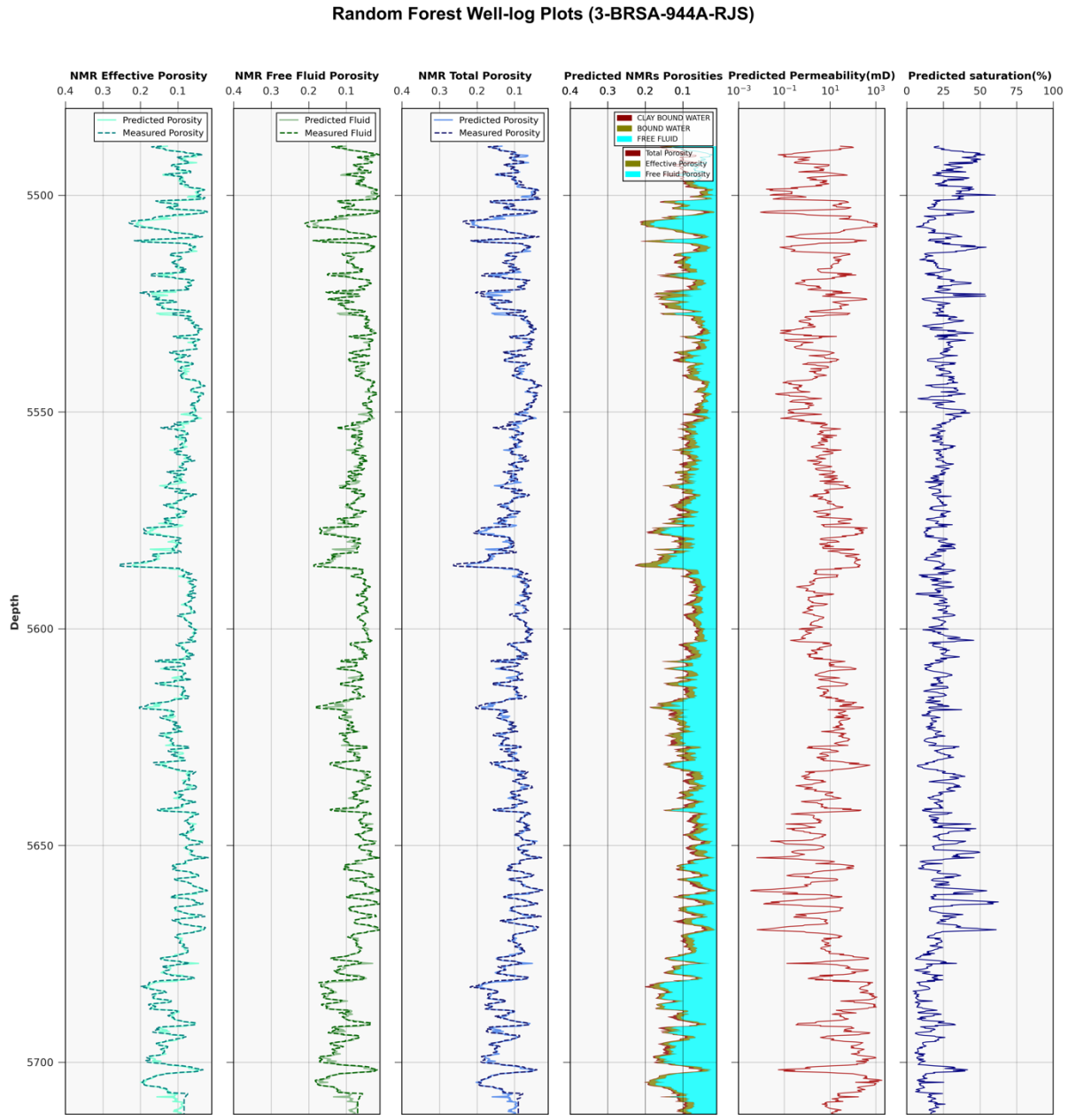

Figure A-11: Comparing the match between the predicted and measured NMR porosity (Effective, Free fluid Total), and Predicted Permeability & Saturation Using Timur-Coates Correlation, for the RF Model, applied to the well (3-BRSA-944A-RJS) included in training phase: (From left to right) 1st track: Predicted and Measured NMR Effective Porosity (m3/m3). 2nd track: Predicted and Measured NMR Free Fluid (m3/m3). 3rd track: Predicted and Measured NMR Total Porosity (m3/m3). 4th track: Predicted Effective, Free Fluid, Total Porosities (m3/m3), 5th track: Timur-Coates Predicted Permeability (mD), 6th track: Timur-Coates Predicted Irreducible water Saturation (%)

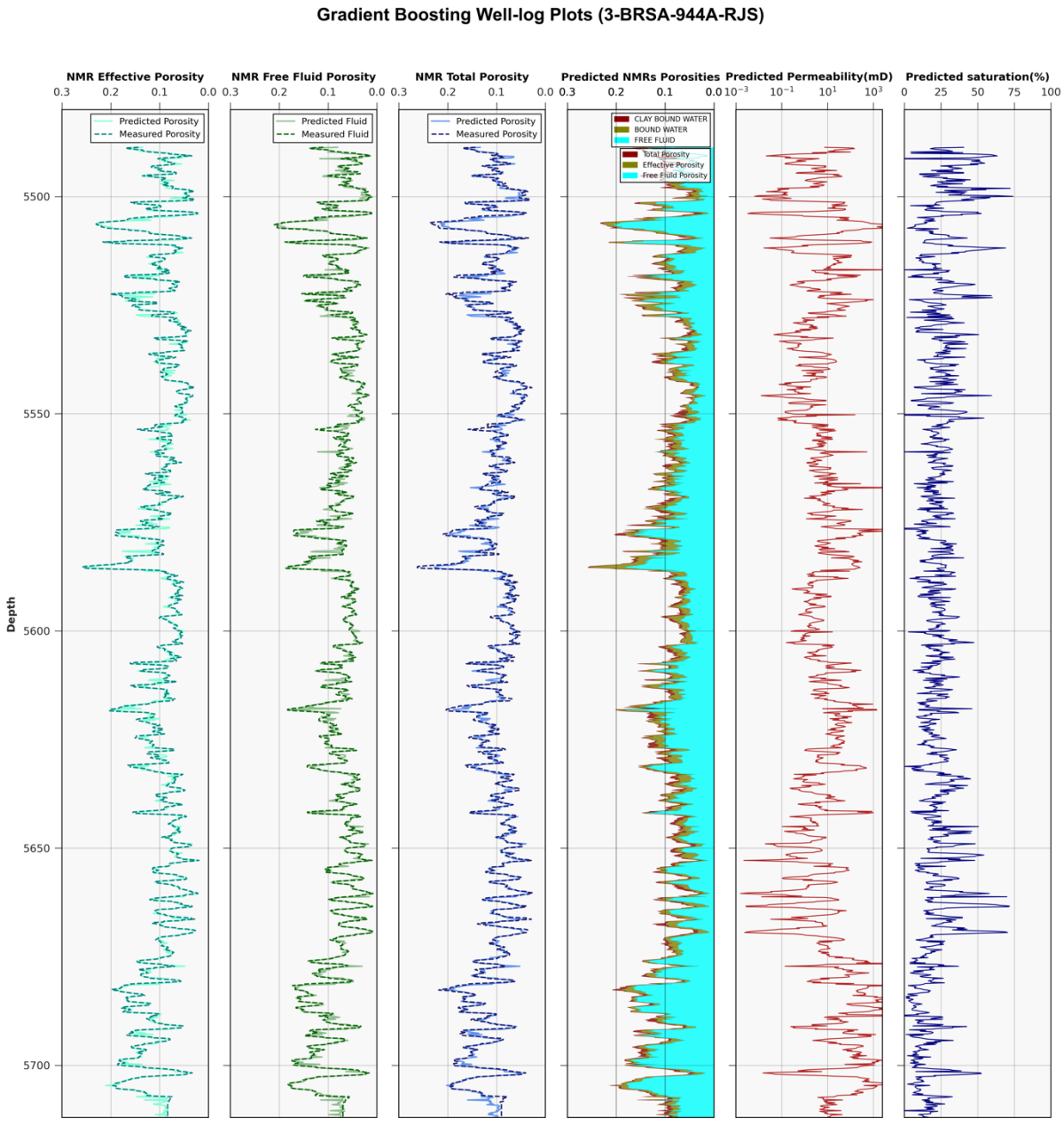

Figure A-12: Comparing the match between the predicted and measured NMR porosity (Effective, Free fluid Total), and Predicted Permeability & Saturation Using Timur-Coates Correlation, for the GB Model, applied to the well (3-BRSA-944A-RJS) included in training phase: (From left to right) 1st track: Predicted and Measured NMR Effective Porosity (m3/m3). 2nd track: Predicted and Measured NMR Free Fluid (m3/m3). 3rd track: Predicted and Measured NMR Total Porosity (m3/m3). 4th track: Predicted Effective, Free Fluid, Total Porosities (m3/m3), 5th track: Timur-Coates Predicted Permeability (mD), 6th track: Timur-Coates Predicted Irreducible water Saturation (%)

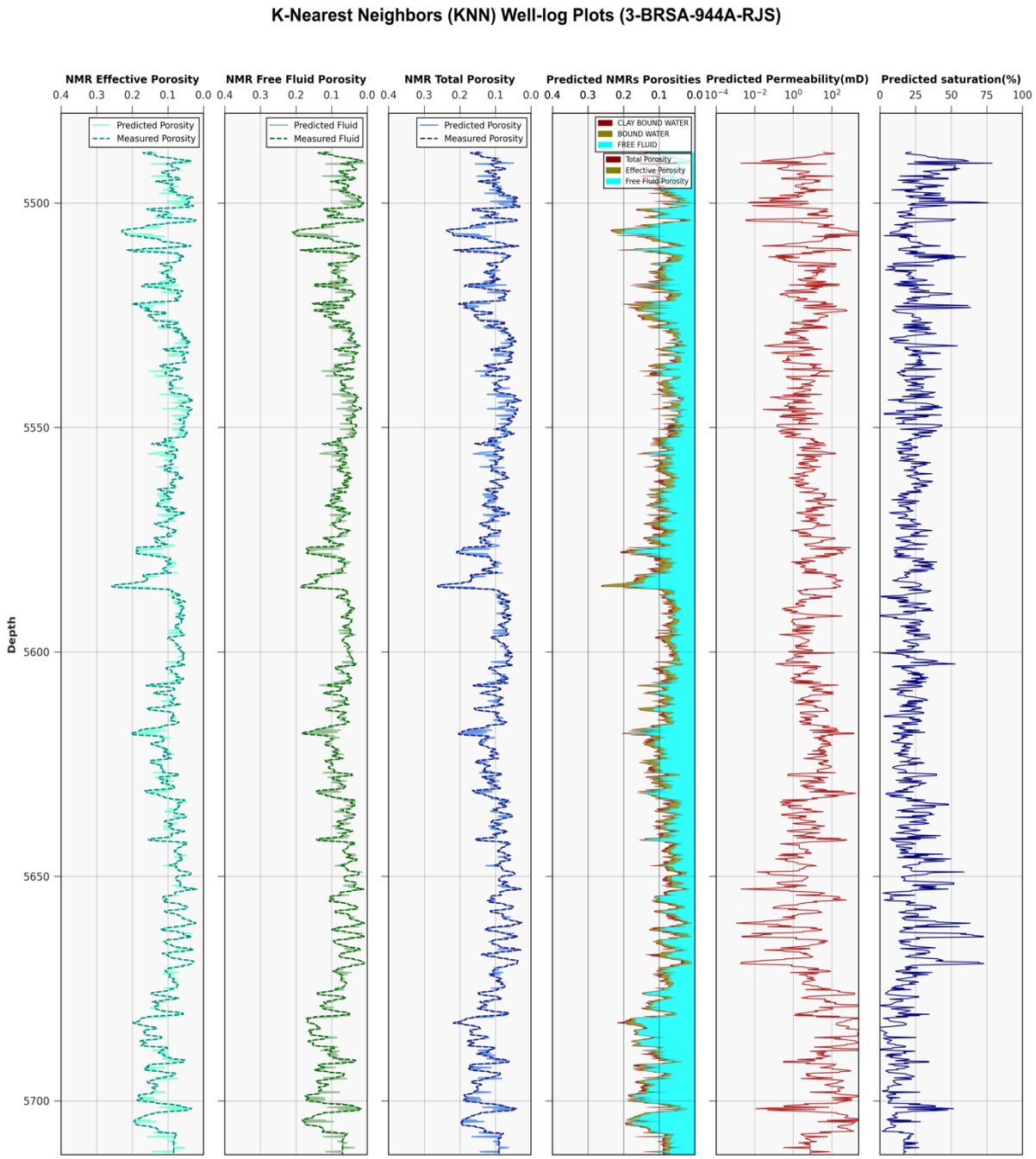

Figure A-13: Comparing the match between the predicted and measured NMR porosity (Effective, Free fluid Total), and Predicted Permeability & Saturation Using Timur-Coates Correlation, for the K-NN Model, applied to the well (3-BRSA-944A-RJS) included in training phase: (From left to right) 1st track: Predicted and Measured NMR Effective Porosity (m3/m3). 2nd track: Predicted and Measured NMR Free Fluid (m3/m3). 3rd track: Predicted and Measured NMR Total Porosity (m3/m3). 4th track: Predicted Effective, Free Fluid, Total Porosities (m3/m3), 5th track: Timur-Coates Predicted Permeability (mD), 6th track: Timur-Coates Predicted Irreducible water Saturation (%)

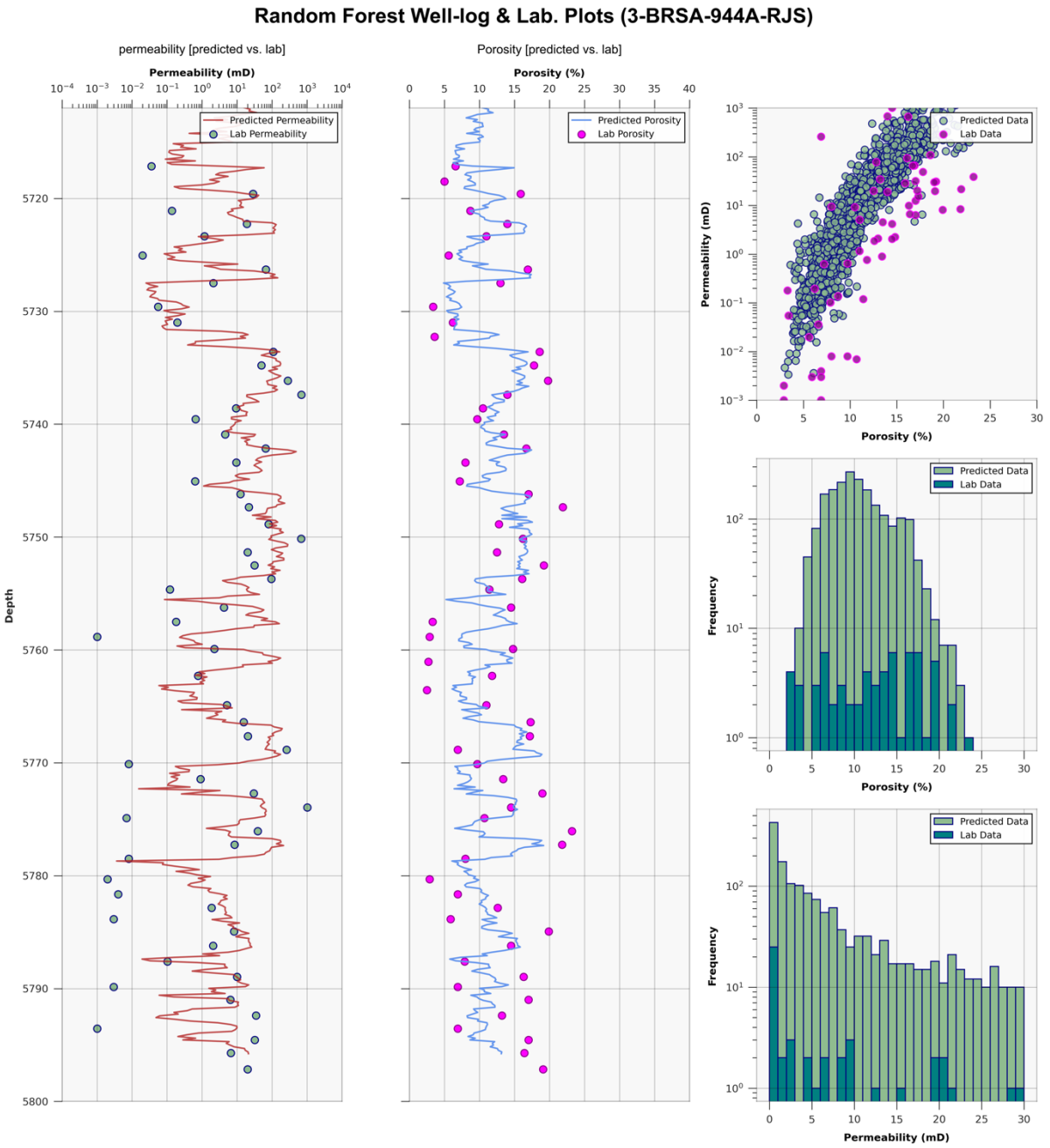

Figure A-14: Comparing the match between the predicted & laboratory measurement of permeability and porosity, Comparing scatter plot permeability-porosity of predicted and laboratory data, Histogram of Porosity and Permeability distribution of predicted and laboratory data, for the RF Model, applied to the well (3-BRSA-944A-RJS) included in training set (for Permeability Prediction Timur-Coates Correlation has been used): (From left to right) 1st track: Predicted and laboratory Permeability (mD). 2nd track: Predicted and Measured Porosity (%). 3rd track: Scatter plot of permeability-porosity of predicted and laboratory data (mD vs. %), 4th track: Histogram of Porosity of predicted and laboratory data (%), 5th track Histogram of Permeability of predicted and laboratory data (mD)

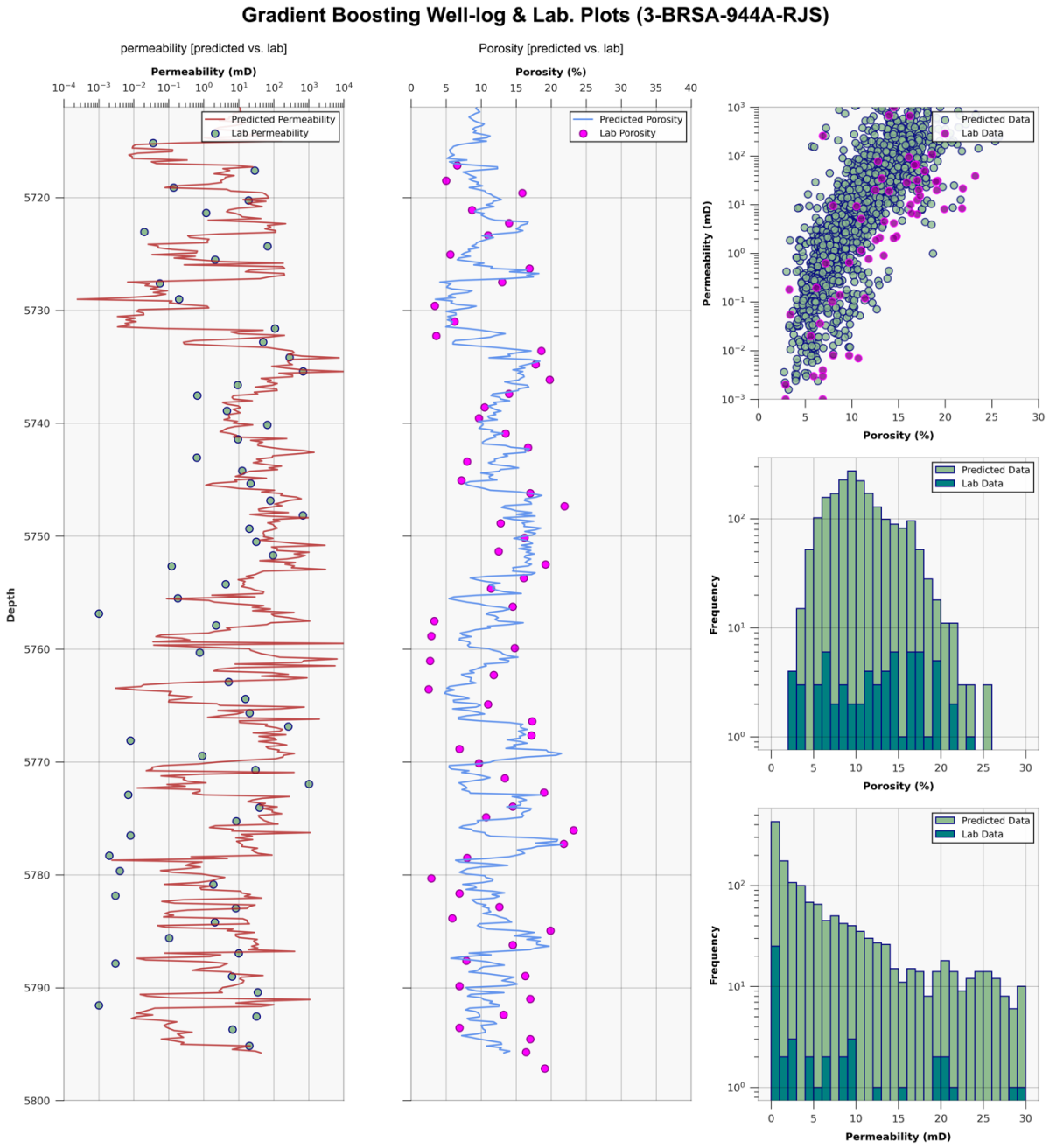

Figure A-15: Comparing the match between the predicted & laboratory measurement of permeability and porosity, Comparing scatter plot permeability-porosity of predicted and laboratory data, Histogram of Porosity and Permeability distribution of predicted and laboratory data, for the GB Model, applied to the well (3-BRSA-944A-RJS) included in training set (for Permeability Prediction Timur-Coates Correlation has been used): (From left to right) 1st track: Predicted and laboratory Permeability (mD). 2nd track: Predicted and Measured Porosity (%). 3rd track: Scatter plot of permeability-porosity of predicted and laboratory data (mD vs. %), 4th track: Histogram of Porosity of predicted and laboratory data (%), 5th track Histogram of Permeability of predicted and laboratory data (mD)

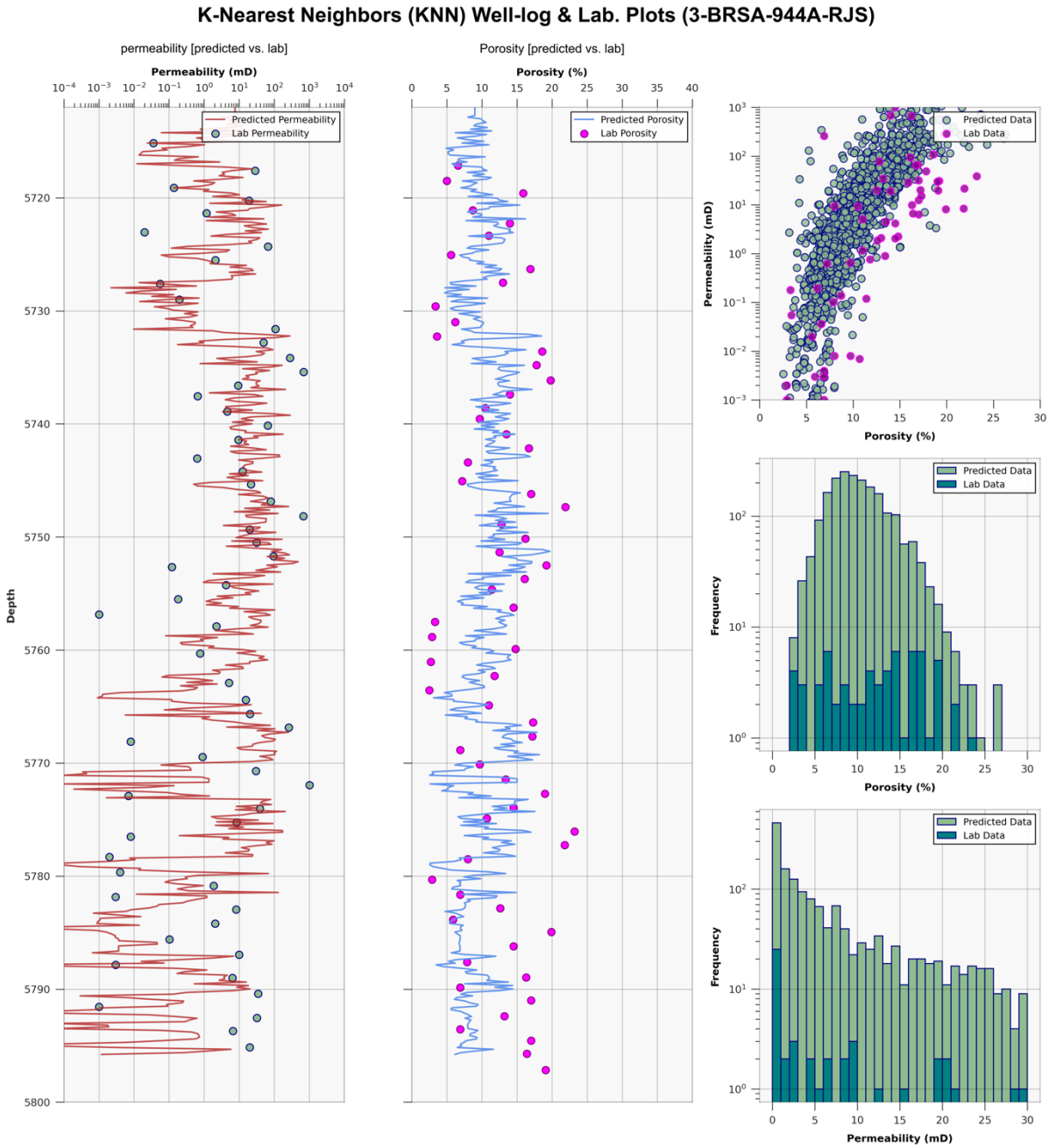

Figure A-16: Comparing the match between the predicted & laboratory measurement of permeability and porosity, Comparing scatter plot permeability-porosity of predicted and laboratory data, Histogram of Porosity and Permeability distribution of predicted and laboratory data, for the K-NN Model, applied to the well (3-BRSA-944A-RJS) included in training set (for Permeability Prediction Timur-Coates Correlation has been used): (From left to right) 1st track: Predicted and laboratory Permeability (mD). 2nd track: Predicted and Measured Porosity (%). 3rd track: Scatter plot of permeability-porosity of predicted and laboratory data (mD vs. %), 4th track: Histogram of Porosity of predicted and laboratory data (%), 5th track Histogram of Permeability of predicted and laboratory data (mD)

Evaluation of different model performances using the different dataset [9-ITP-RJS] (Not included in training-test)

Scatter plots

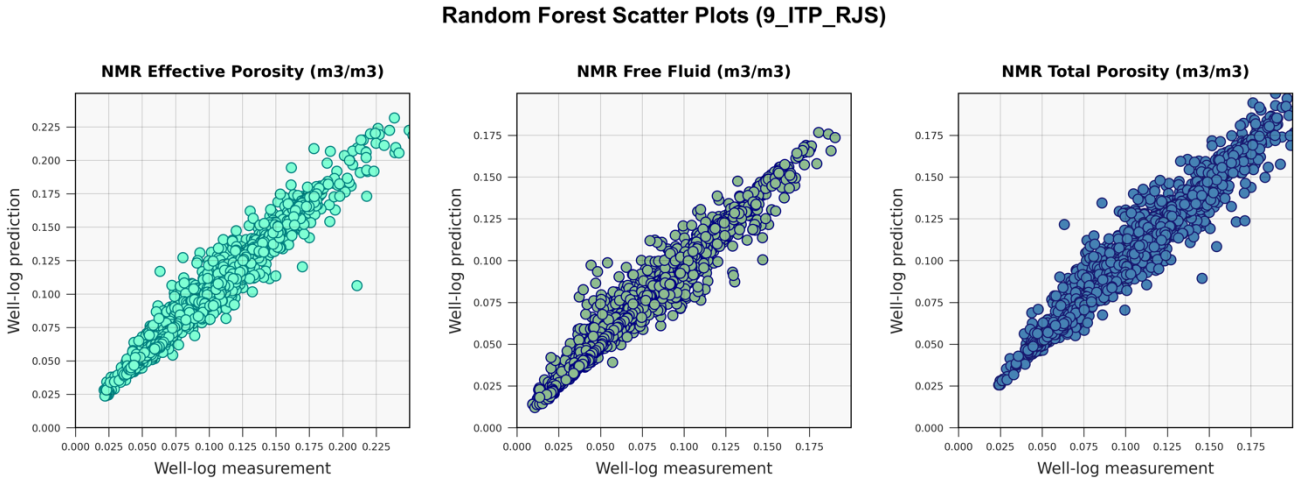

Figure A-17: Scatter plots of predicted versus measured NMR porosity, for the RF Model, applied to one of the well (9-ITP-RJS) included in training phase: Left plot: NMR Effective Porosity (m3/m3). Middle plot: NMR Free Fluid (m3/m3). Right plot: NMR Total Porosity (m3/m3).

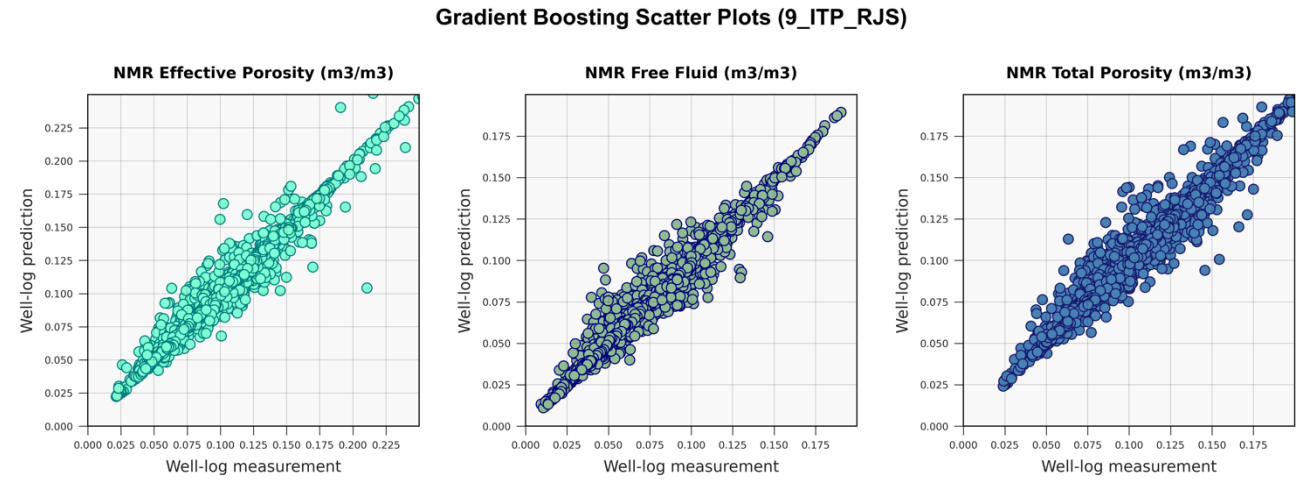

Figure A-18: Scatter plots of predicted versus measured NMR porosity, for the GB Model, applied to one of the well (9-ITP-RJS) included in training phase: Left plot: NMR Effective Porosity (m3/m3). Middle plot: NMR Free Fluid (m3/m3). Right plot: NMR Total Porosity (m3/m3).

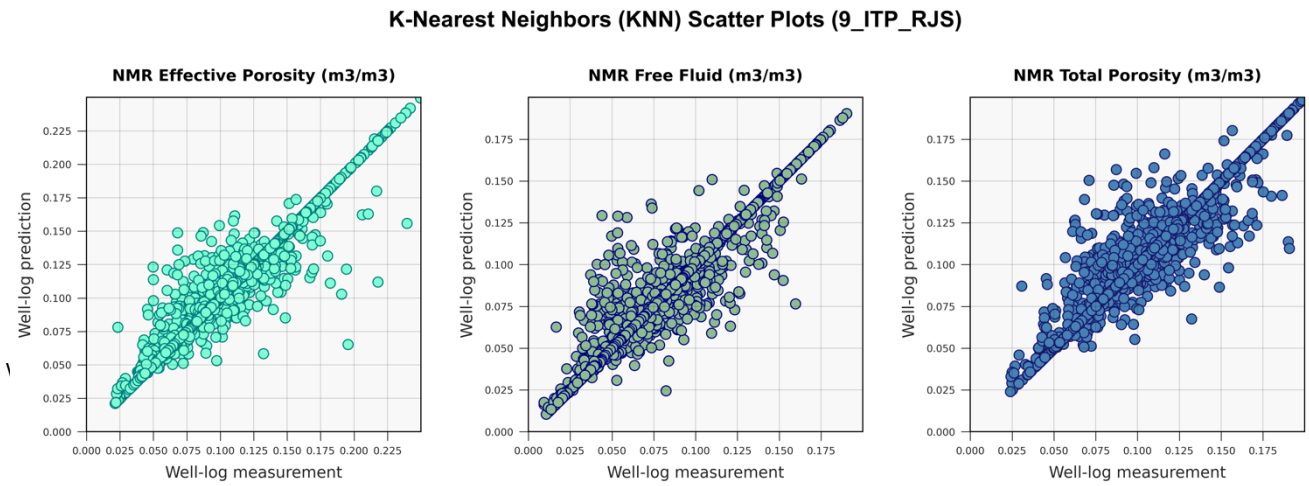

Figure A-19: Scatter plots of predicted versus measured NMR porosity, for the K-NN Model, applied to one of the well (9-ITP-RJS) included in training phase: Left plot: NMR Effective Porosity (m3/m3). Middle plot: NMR Free Fluid (m3/m3). Right plot: NMR Total Porosity (m3/m3).

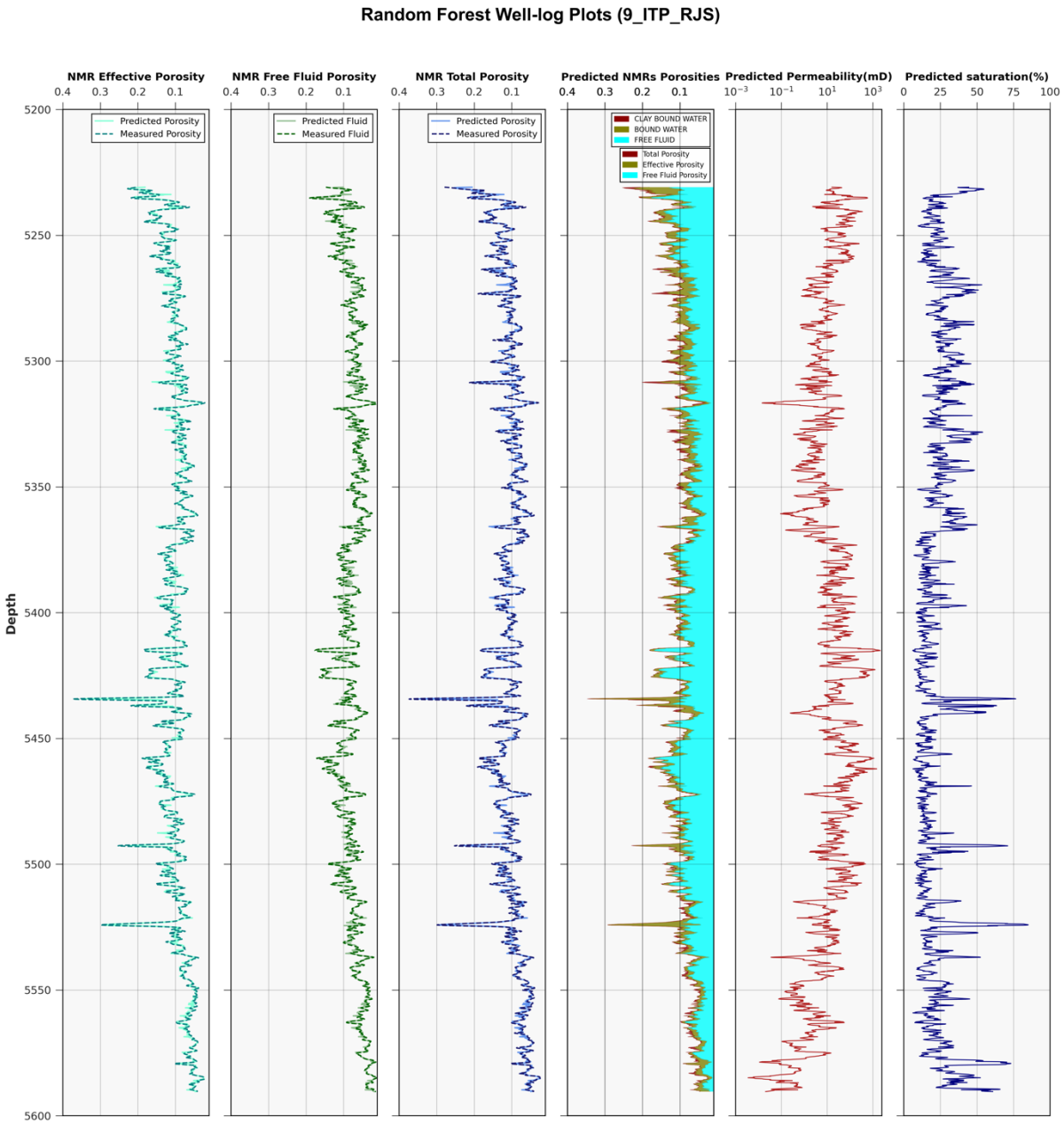

Figure A-20: Comparing the match between the predicted and measured NMR porosity (Effective, Free fluid Total), and Predicted Permeability & Saturation Using Timur-Coates Correlation, for the RF Model, applied to the well (9-ITP-RJS) included in training phase: (From left to right) 1st track: Predicted and Measured NMR Effective Porosity (m3/m3). 2nd track: Predicted and Measured NMR Free Fluid (m3/m3). 3rd track: Predicted and Measured NMR Total Porosity (m3/m3). 4th track: Predicted Effective, Free Fluid, Total Porosities (m3/m3), 5th track: Timur-Coates Predicted Permeability (mD), 6th track: Timur-Coates Predicted Irreducible water Saturation (%)

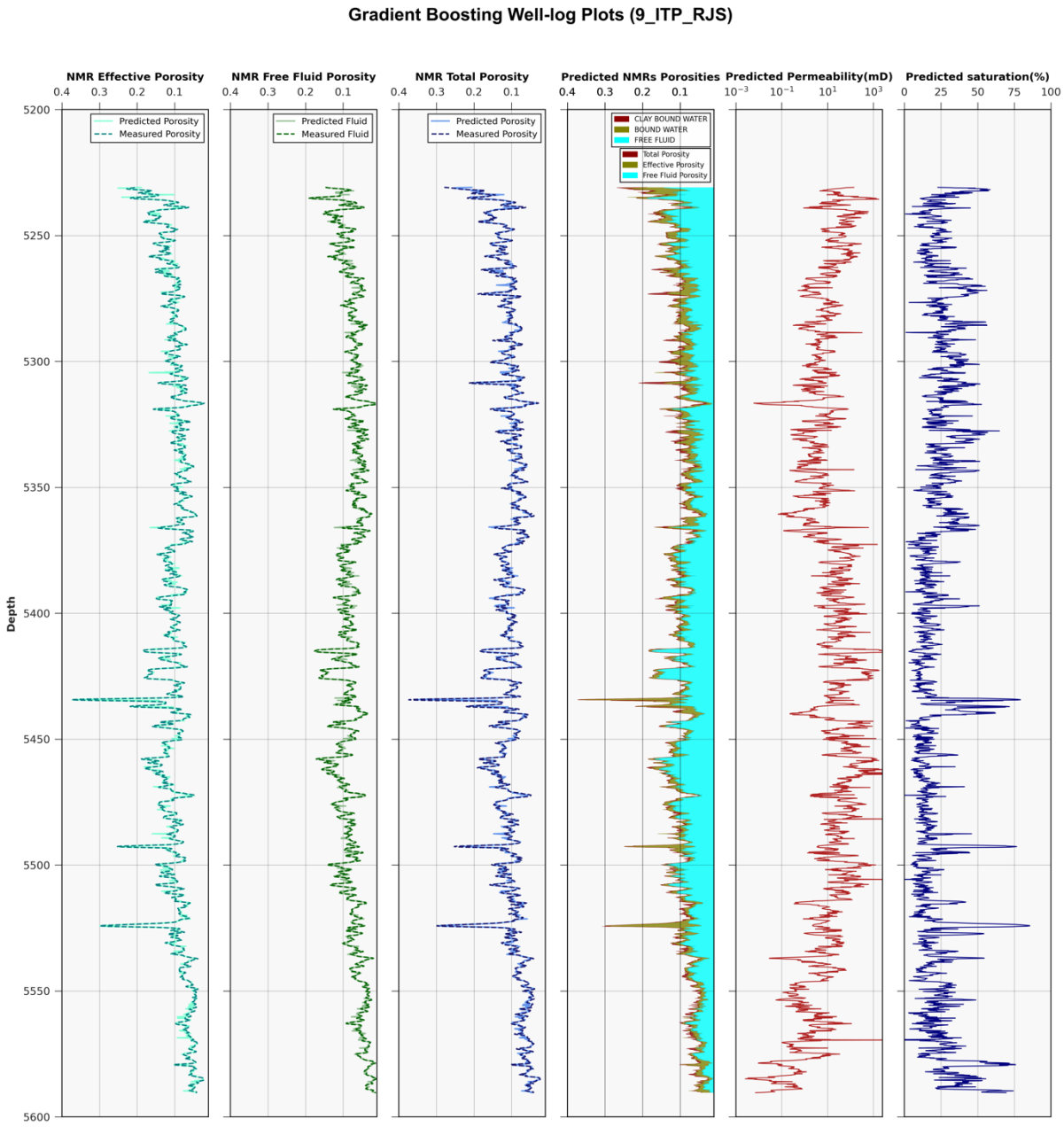

Figure A-21: Comparing the match between the predicted and measured NMR porosity (Effective, Free fluid Total), and Predicted Permeability & Saturation Using Timur-Coates Correlation, for the GB Model, applied to the well (9-ITP-RJS) included in training phase: (From left to right) 1st track: Predicted and Measured NMR Effective Porosity (m3/m3). 2nd track: Predicted and Measured NMR Free Fluid (m3/m3). 3rd track: Predicted and Measured NMR Total Porosity (m3/m3). 4th track: Predicted Effective, Free Fluid, Total Porosities (m3/m3), 5th track: Timur-Coates Predicted Permeability (mD), 6th track: Timur-Coates Predicted Irreducible water Saturation (%)

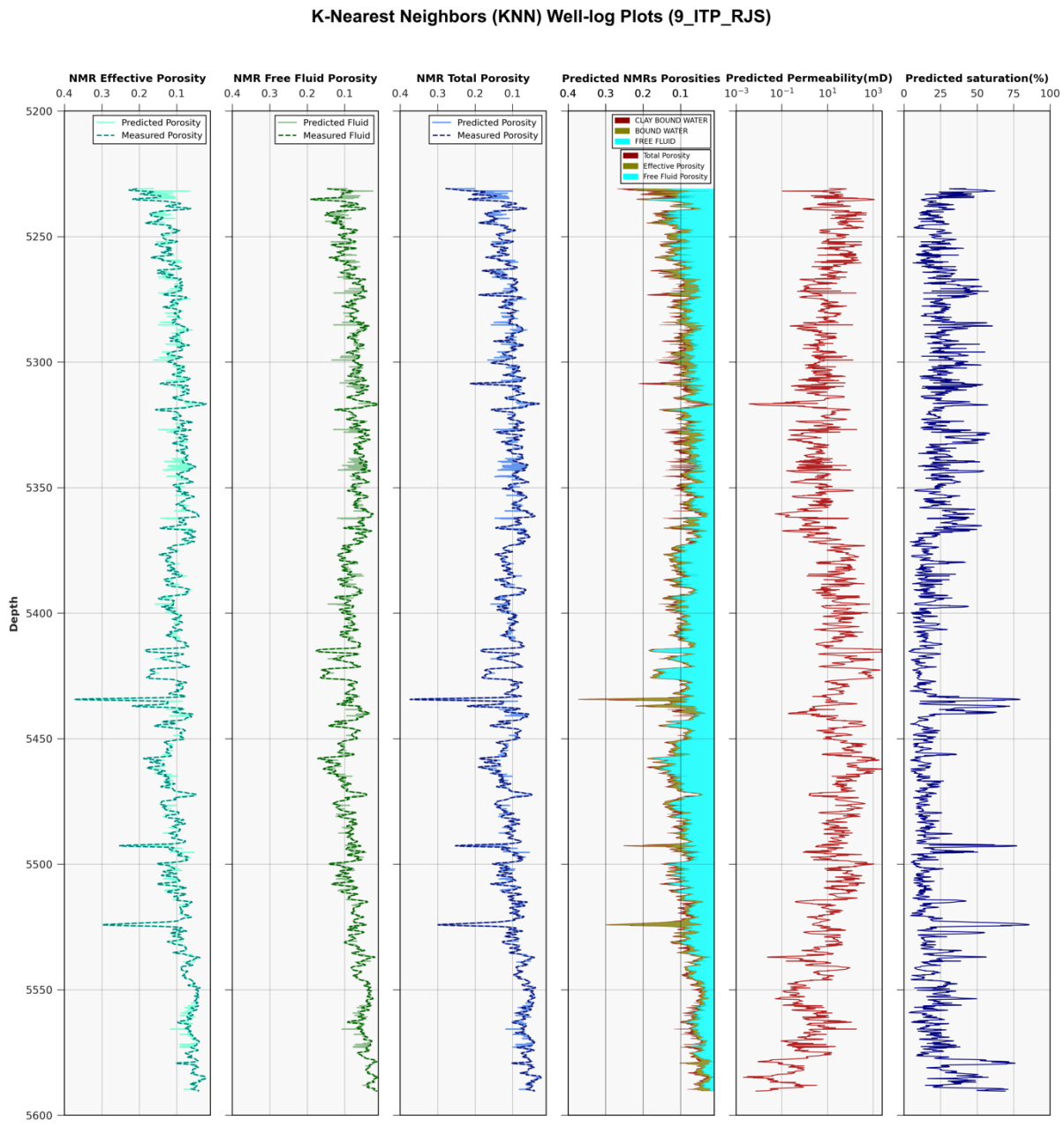

Figure A-22: Comparing the match between the predicted and measured NMR porosity (Effective, Free fluid Total), and Predicted Permeability & Saturation Using Timur-Coates Correlation, for the K-NN Model, applied to the well (9-ITP-RJS) included in training phase: (From left to right) 1st track: Predicted and Measured NMR Effective Porosity (m3/m3). 2nd track: Predicted and Measured NMR Free Fluid (m3/m3). 3rd track: Predicted and Measured NMR Total Porosity (m3/m3). 4th track: Predicted Effective, Free Fluid, Total Porosities (m3/m3), 5th track: Timur-Coates Predicted Permeability (mD), 6th track: Timur-Coates Predicted Irreducible water Saturation (%)

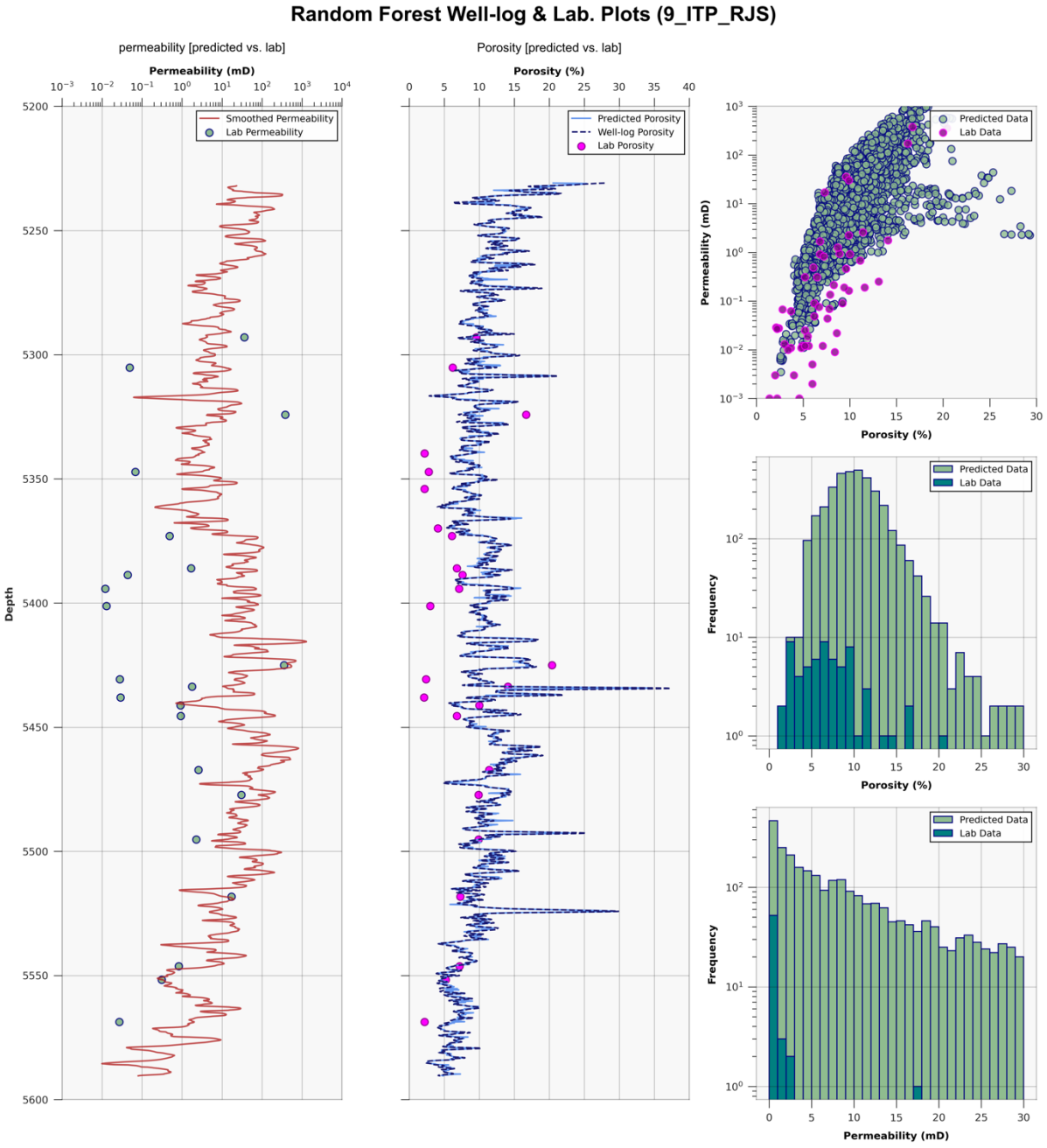

Figure A-23: Comparing the match between the predicted & laboratory measurement of permeability and porosity, Comparing scatter plot permeability-porosity of predicted and laboratory data, Histogram of Porosity and Permeability distribution of predicted and laboratory data, for the RF Model, applied to the well (3-BRSA-944A-RJS) included in training set (for Permeability Prediction Timur-Coates Correlation has been used): (From left to right) 1st track: Predicted and laboratory Permeability (mD). 2nd track: Predicted and Measured Porosity (%). 3rd track: Scatter plot of permeability-porosity of predicted and laboratory data (mD vs. %), 4th track: Histogram of Porosity of predicted and laboratory data (%), 5th track Histogram of Permeability of predicted and laboratory data (mD)

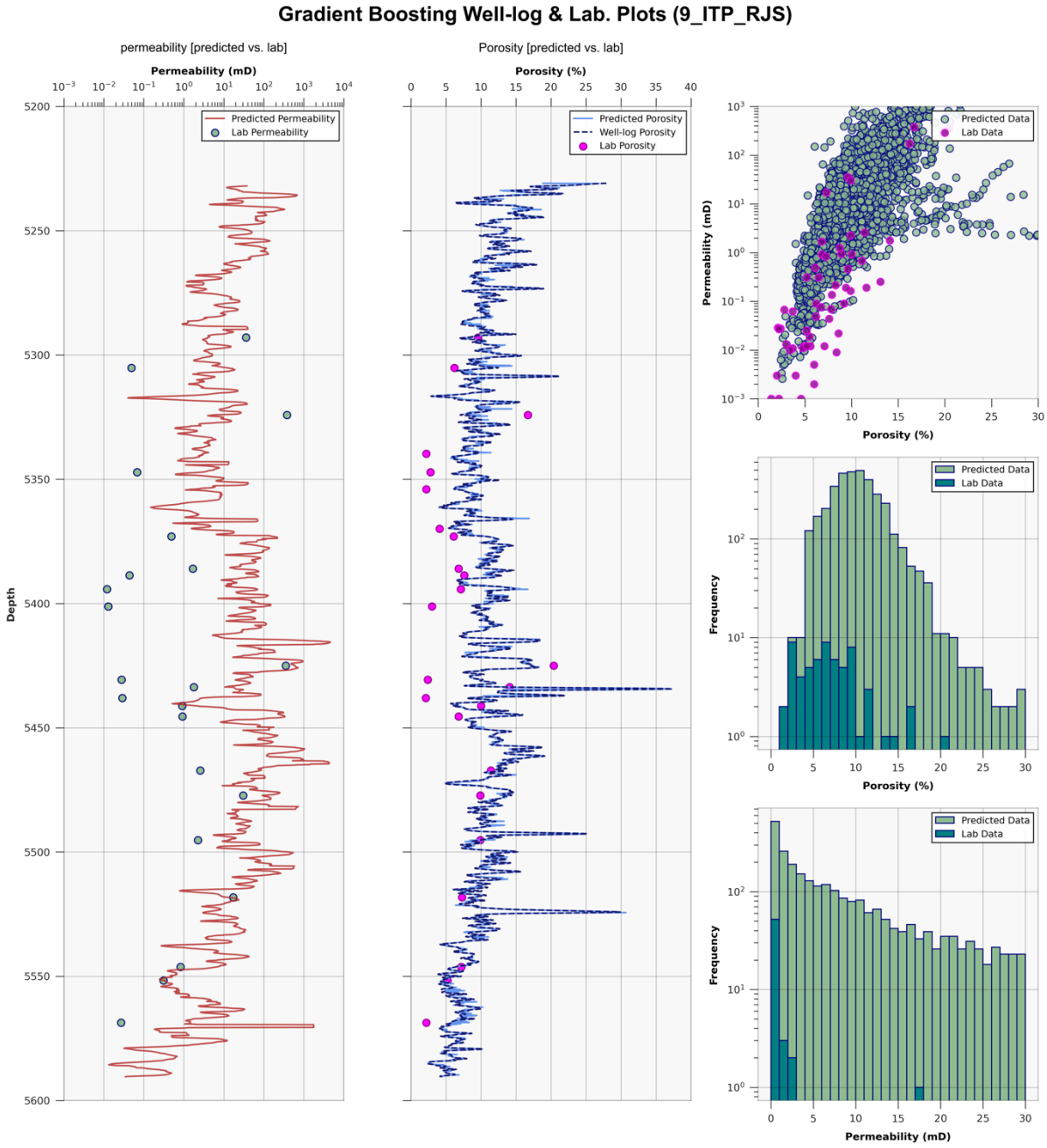

Figure A-24: Comparing the match between the predicted & laboratory measurement of permeability and porosity, Comparing scatter plot permeability-porosity of predicted and laboratory data, Histogram of Porosity and Permeability distribution of predicted and laboratory data, for the GB Model, applied to the well (3-BRSA-944A-RJS) included in training set (for Permeability Prediction Timur-Coates Correlation has been used): (From left to right) 1st track: Predicted and laboratory Permeability (mD). 2nd track: Predicted and Measured Porosity (%). 3rd track: Scatter plot of permeability-porosity of predicted and laboratory data (mD vs. %), 4th track: Histogram of Porosity of predicted and laboratory data (%), 5th track Histogram of Permeability of predicted and laboratory data (mD)

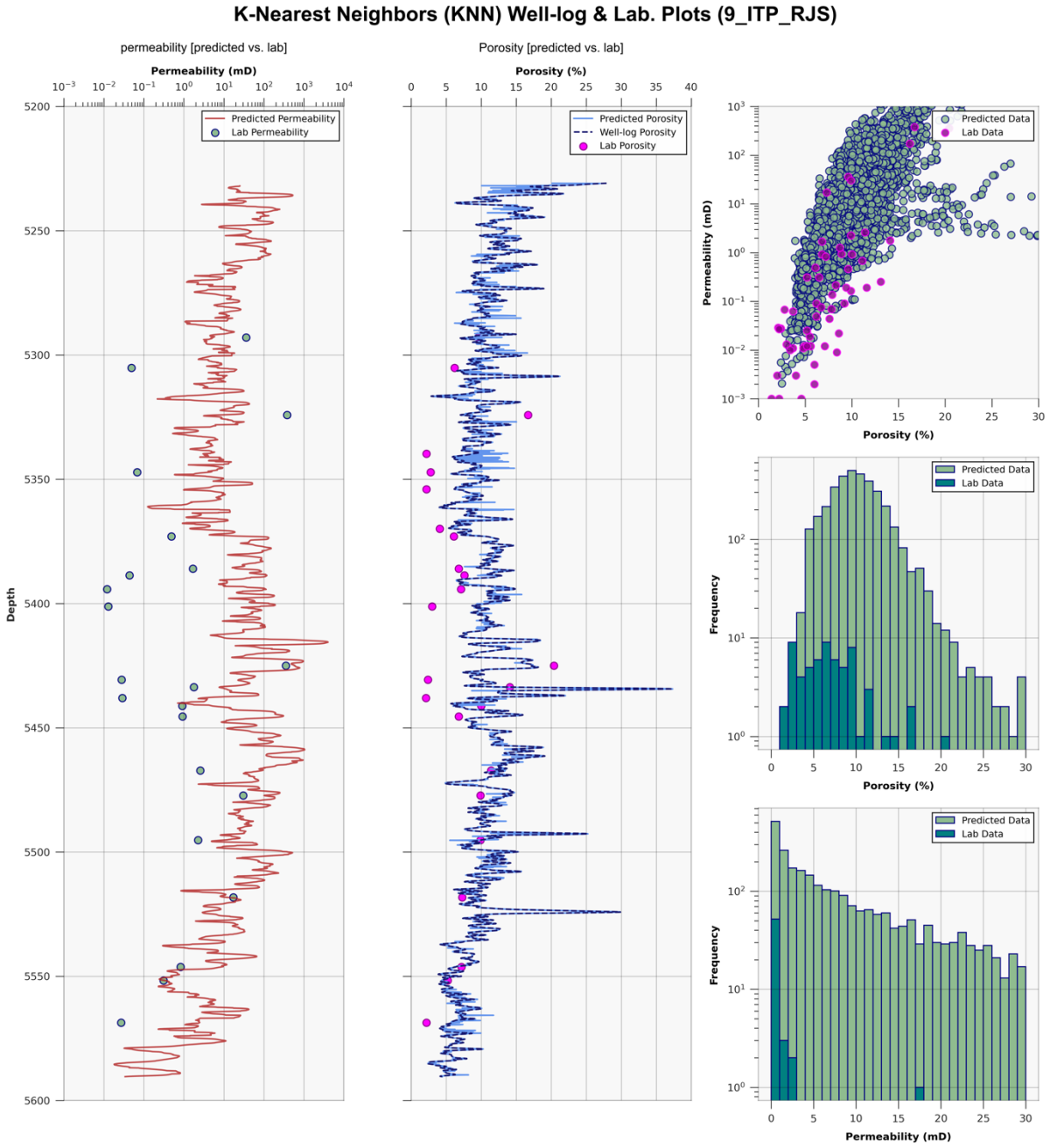

Figure A-25: Comparing the match between the predicted & laboratory measurement of permeability and porosity, Comparing scatter plot permeability-porosity of predicted and laboratory data, Histogram of Porosity and Permeability distribution of predicted and laboratory data, for the K-NN Model, applied to the well (3-BRSA-944A-RJS) included in training set (for Permeability Prediction Timur-Coates Correlation has been used): (From left to right) 1st track: Predicted and laboratory Permeability (mD). 2nd track: Predicted and Measured Porosity (%). 3rd track: Scatter plot of permeability-porosity of predicted and laboratory data (mD vs. %), 4th track: Histogram of Porosity of predicted and laboratory data (%), 5th track Histogram of Permeability of predicted and laboratory data (mD)

Appendix 03: (Machine Learning Model approaches)

Support Vector Machine (SVM)

SVM was first proposed by Vapnik in and has been successfully applied to regression problems including grid load forecasting, fault diagnosis, and image processing. SVM shows great performance in high-dimensional function approximation problems due to the use of the kernel technique, which maps feature vectors to a higher-dimensional space. It is one of the most popular and versatile models in ML, suitable for both classification and regression of complex small datasets. Hence, many researchers use SVM to estimate the SOH of batteries. The architecture of the SVM method for regression is shown in Fig. 10. In general, the SVM model is defined as:

$$\hat{y} = w^T.\psi(x) + b, \qquad x \in R^d, \qquad \psi(x) \in R^{\tilde{d}}, \qquad b \in R$$

Equation A-1

where $\psi(.)$ is a mapping that makes the input data linear in a new feature space with dimension $\tilde{d}$. Different from the general linear regression models, the SVM model uses the ε-insensitive loss function. This states that any error larger than ε is deemed unacceptable. That is, the objective of the basic SVM is to find the optimal coefficients w and b such that the function, f, does not contain errors larger than ε. This is, therefore, also called the hard-margin SVM. The hard-margin SVM leads to the following constrained optimization problem.

$$\min \frac{1}{2} w^T w$$

$$s.t. \begin{cases} y_i - w^T.\psi(x_i) - b \le \varepsilon \\ w^T.\psi(x_i) + b - y_i \le \varepsilon \end{cases}, \qquad \forall i \in \{1, 2, \dots, N\}$$

Equation A-2

However, it is not always feasible to find a minimum under these constraints. Therefore, the following loss function is introduced:

$$\xi_{\epsilon}(\hat{y}_i, y_i) = \begin{cases} 0 & , |y_i - \hat{y}_i| - \epsilon \\ |y_i - \hat{y}_i| - \epsilon & , Otherwize \end{cases}, \qquad \forall i \in \{1, 2, \dots, N\}$$

Equation A-3

Based on Equation 3, the samples with the predicted error less than ε are deemed acceptable, while the samples outside of the ε band will increase the regression error. Slack variables ξ_i and ξ_i^* are introduced to create a soft-margin and allowing for measurement errors, making the optimization feasible with otherwise infeasible constraints. The primal SVM optimization problem has the following form:

$$\min$$

$$w \in R^{\tilde{d}}$$

$$\xi_i, {\xi_i}^* \in R^N$$

$$\frac{1}{2} w^T w + C \sum_i^N (\xi_i + {\xi_i}^*) \; s.t. \; \begin{cases} y_i - w^T.\psi(x_i) - b \leq \epsilon + \xi_i \\ w^T.\psi(x_i) + b - y_i \leq \epsilon + {\xi_i}^* \\ \xi_i \, , {\xi_i}^* \geq 0 \end{cases}, \forall i \in \{1, 2, \dots, N\}$$

Equation A-4

where C is a positive constant regulating the penalty, it determines the trade-off between the flatness of the regression function and the amount to which deviations larger than ε are tolerated. Flatness in the case of equation 4 means a small $\|w\|$.In order to solve this problem, the Lagrange multipliers $\alpha_i, {\alpha_i}^*, \beta_i, {\beta_i}^* \geq 0$ are introduced, and the Lagrangian can be expressed as follows:

$$\min$$

$$w \in R^{\tilde{d}}$$

$$\xi_i, {\xi_i}^* \in R^N$$

$$\max_{\alpha, \beta \in [0+\infty]^N} \; L(w, b, \xi_i \, , {\xi_i}^*, \alpha_i, {\alpha_i}^*, \beta_i, {\beta_i}^*)$$

$$= \frac{1}{2} w^T w + C \sum_{i=1}^{N} (\xi_i + {\xi_i}^*) - \sum_{i=1}^{N} \left(\beta_i \xi_i + {\beta_i}^* {\xi_i}^*\right)$$

$$- \sum_{i=1}^{N} \alpha_i \, (\epsilon + \xi_i - y_i + w^T.\psi(x_i) + b) - \sum_{i=1}^{N} {\alpha_i}^* \, (\epsilon + {\xi_i}^* + y_i - w^T.\psi(x_i) - b)$$

Equation A-5

The min–max problem can be transferred into its dual max–min problem which satisfies the Karushe-Kuhne-Tucker (KKT) conditions. The first KKT condition states that the gradients of the primal variables are equal to zero i.e., $\nabla_w L = 0$, $\nabla_b L = 0$, $\nabla_{\xi_i} L = 0$, $\nabla_{{\xi_i}^*} L = 0$.

The second KKT condition called the complementary conditions states that multiplying the constraint by its Lagrange multiplier has to equal zero in the optimum. That is, either the constraint is active, or the Lagrange multiplier is zero. As a consequence of the second KKT condition, the Lagrange multiplier α_i and ${\alpha_i}^*$ for the samples inside the ε–tube will vanish; while when $|y_i - \hat{y}_i| \geq \varepsilon$, the multipliers α_i and ${\alpha_i}^*$ are nonzero. Therefore, only the samples x_i with non-vanishing coefficients are enough to describe w, and these samples are commonly called the support vectors (SVs). The primal SVM optimization problem is converted into the following dual SVM optimization problem

$$\max_{\alpha,\alpha^*} \sum_{i=1}^{N} y_i(\alpha_i^* - \alpha_i) - \sum_{i=1}^{N} \epsilon(\alpha_i^* + \alpha_i) - \frac{1}{2}\sum_{i=1}^{N}\sum_{i=1}^{N}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)K(x_i, x_j) \text{ s.t.} \begin{cases} \sum_{i=1}^{N}(\alpha_i^* - \alpha_i) = 0 \\ 0 \le \alpha_i, \quad \alpha_i^* \le C \end{cases}$$

Equation A-6

After optimizing equation 6, w.r.t. the Lagrange multipliers α_i and α_i^*, the coefficients w and b can be computed from the α's using equation 7 and 8, respectively.

$$w = \sum_{i=1}^{N}(\alpha_i - \alpha_i^*).\psi(x_i)$$

Equation A-7

$$b = y_i - \sum_{i=1}^{N}(\alpha_i - \alpha_i^*)\psi(x_i)^T.\psi(x_i), \qquad \text{for example i where } 0 < \alpha_i^*, \alpha_i < C$$

Equation A-8

Finally, the regression function can be described as:

$$f(x) = w^T.\psi(x_i) + b = \sum_{i=1}^{N}(\alpha_i^* - \alpha_i)^T.K(x_i, x_j)$$

Equation A-9

Where $K(x_i, x) = \langle \psi(x_i), \psi(x) \rangle$ is the kernel function. The kernel function implicitly maps the input to the high-dimensional feature space. This method has higher computational efficiency than if the features were first mapped using $\psi(\cdot)$, thereby, overcoming the curse of dimensionality. Common kernel functions $K(x_{ii}, x_{jj})$ used in SVM are:

- Polynomial kernel:

$$K(x_{ii}, x_{jj}) = ({x_{ii}}^T.x_{jj} + 1)^M$$

Equation A-10

- Gaussian radial basis function:

$$K(x_{ii}, x_{jj}) = \left(-\frac{1}{2\sigma^2}\|x_{ii} - x_{jj}\|^2\right)$$

Equation A-11

- Hyperbolic tangent kernel:

The Hyperbolic tangent kernel often used as an activation function for artificial neurons, expressed as:

$$K(x_{ii}, x_{jj}) = tanh(\kappa {x_{ii}}^T . x_{jj} + c)$$

Equation A-12

where M, σ, κ, c are adjustable parameters of the above kernel functions.

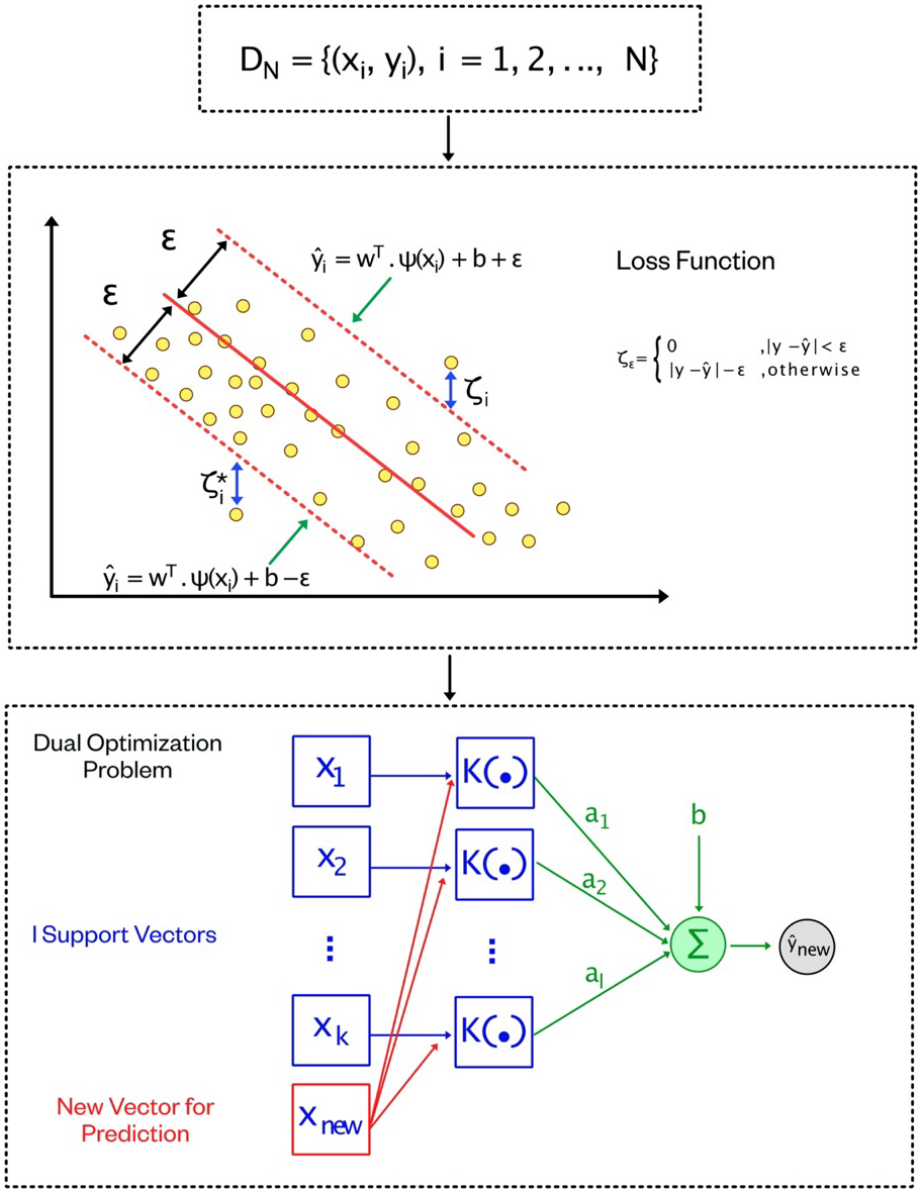

Figure A-26: The illustration of Support Vector Machine (SVM).

Compared with the optimization problem of the standard SVM, given in equation 4, LS-SVM has a less computational burden and faster solving speed because it solves linear equations instead of quadratic

programming problems. Before the model training, the kernel principal components analysis (PCA) algorithm was introduced to fuse these features. As a result, the obtained self-adaptive feature shows higher relevance to the battery capacity than most of the single features. Also, the LS-SVM model is optimized by the particle swarm optimization (PSO) algorithm.

K-Nearest Neighbour Regression (KNN)

K-NN is efficient for classification purposes in pattern recognition. As a kind of lazy learning, k-NN uses the k closest neighbors in the feature space to classify a new point. When used for regression, as presented in Fig. 11, k-NN first finds the k closest points x^1, x^2, …, x^k of a new point x_{new} based on a distance measure, and calculates the weighted average of their response to predict the response of x_{new}. For a given training dataset with N points $\mathrm{X} = \{x_1, x_2, \dots, x_N\}$, where each point possessesd features, the response of a new data x_{new} can be estimated by k-NN as follows.

First, in order to describe how close each training points x_1 is to the testing points x_{new}, the weighted Euclidean distance between them is calculated, which can be expressed as

$$d(x_i, x_{new}) = \sqrt{\sum_{j=1}^{d} w_j ({x_{new}}^j - {x_i}^j)^2}$$

Equation A-13

Where $x_{new,j}$ and $x_{i,j}$ are the j_{th} feature of the new point x_{new} and the training points x_i, respectively. Besides, wj is the weight of j_{th} feature, with the weights being subjected to the constraint $\sum_{j=1}^{d} w_j = 1$. The weight w_j reflects the importance of the feature and can be found using an optimization algorithm, such as PSO, or differential evolution (DE) algorithm. According to the distance d, the k training points $Z = \{x_{(1)}, x_{(2)}, \dots, x_{(N)}\}$ ordered from the nearest to furthest are obtained. These are called the k nearest neighbour of x_{new}. A kernel function is then used to assign weights to each neighbour (the kernel is usually dependent on the calculated distance), and the prediction for new sample x_{new} can be obtained by:

$$\hat{y}_{new} = \frac{\sum_{i=1}^{k} K(x_{new}, x_{(i)}) y_i}{\sum_{i=1}^{k} K(x_{new}, x_{(i)})}$$

Equation A-14

where κ represents the number of nearest neighbors which controls the flexibility of the model (the higher k is the smoother the model is going to be). $y_{(i)}$ represents the known response of $y_{(i)}$, $\hat{y}_{new}$ the predicted response of x_{new}, and $K(x_{new}, x_{(i)})$ denotes the kernel function, as given in equations 10-12. The principle of k-NN regression is simple and is easy to be implemented. The PSO algorithm obtained the optimal combination of the feature weight. It not only shows the relative importance of

each feature but also ensures accurate capacity estimation. Even though the k-NN regression model is simple and an accurate SOH estimation is easily obtained, the algorithm has a clear disadvantage: the entire range of the battery degradation has to be known, as the k-NN model cannot predict values outside of the observed range.

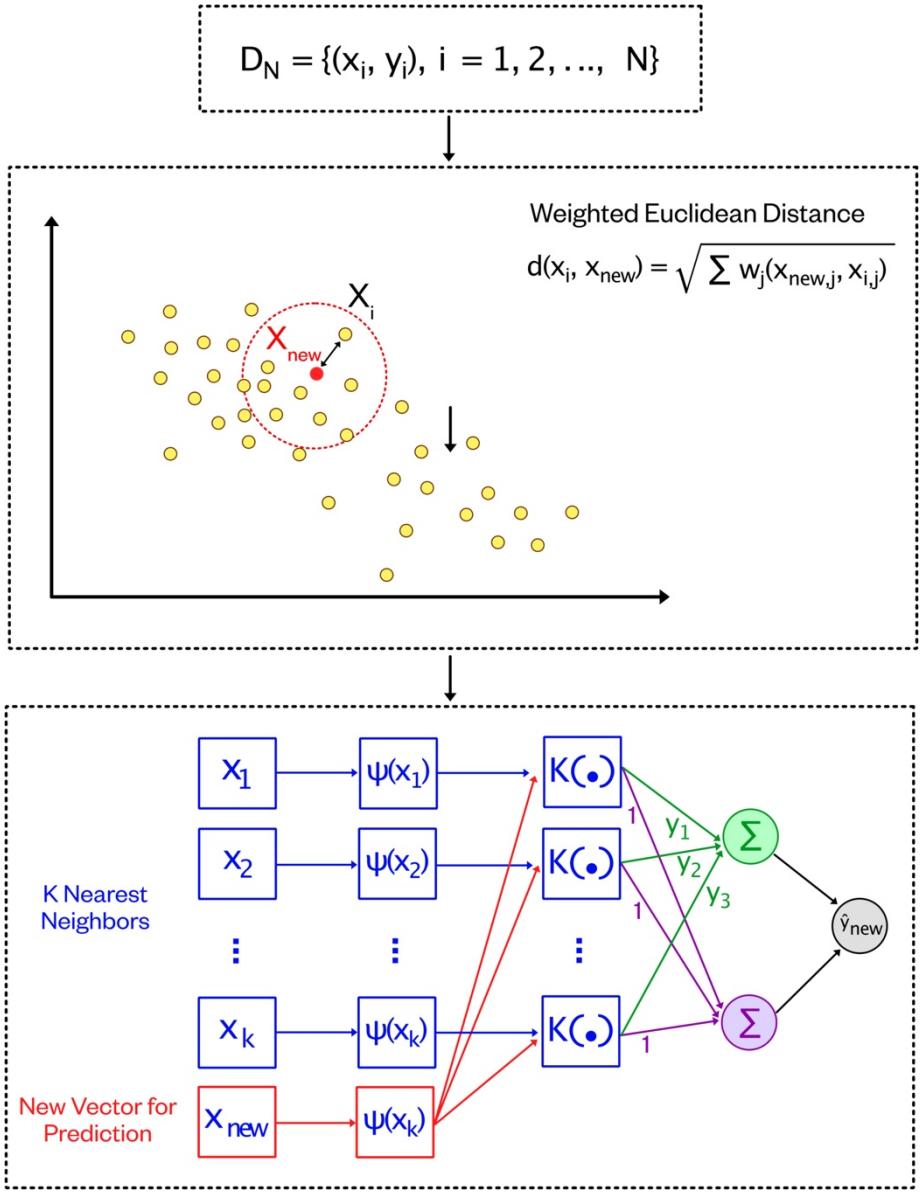

Figure A-27: The illustration of K-Nearest Neighbor Regressor (K-NN).

Appendix 04: (Developing Machine Learning Model)

We have included specific lines of Python code that were written for predicting NMR porosity logs and permeability from well-log data.

The complete Python code can be accessed through the following link (GitHub Repository):

Link to Git HUB

References

Lee, S., Kharghoria, A., & Datta-Gupta, A. 2002. *Electrofacies characterization and permeability predictions in complex reservoirs.* s.l. : Spe Reservoir Evaluation & Engineering, 2002.

Bueno, J., Honório, B., Kuroda, M., Vidal, A., & Leite, E. 2014. *Structural and stratigraphic feature delineation and facies distribution using seismic attributes and well log analysis applied to a brazilian carbonate field.* 2014.

Al-Mudhafar, Watheq J. 2007. *Integrating well log interpretations for lithofacies classification and permeability modeling through advanced machine learning algorithms.* s.l. : Journal of Petroleum Exploration and Production Technology, 2007.

Mihály Dobróka, Norbert Péter Szabó, József Tóth, and Péter Vass. 2016. *Interval inversion approach for an improved interpretation of well logs.* s.l. : Society of Exploration Geophysicists, 2016.

Aleksandr Kochnev, Sergey Galkin, Sergey Krivoshchekov, Nikita Kozyrev, Polina Chalova. 2021. *Application of Machine Learning Algorithms to Predict the Effectiveness of Radial Jet Drilling Technology in Various Geological Conditions.* s.l. : Applied Science, 2021.

Zhao Wang, Hongming Tang, Yawei Hou, Hongfu Shi, Jingling Li, Tao Yang, Yutian Feng, Wang Meng. 2022. *Quantitative evaluation of unconsolidated sandstone heavy oil reservoirs based on machine learning.* s.l. : geological Journal, 2022.

—. 2022. *Quantitative evaluation of unconsolidated sandstone heavy oil reservoirs based on machine learning.* s.l. : Geological Journal, 2022.

—. 2022. *Quantitative evaluation of unconsolidated sandstone heavy oil reservoirs based on machine learning.* s.l. : Geological Journal, 2022.

Pellegrini, Vittoria De. 2023. *Development of Supervised Machine Learning Models for the Prediction of Well-Logs & Application on Wells at Sao Francisco and Santos Basins, Brazil.* Turin : Politecnico di Torino, 2023.

Rezaee, . 2022. *Synthesizing Nuclear Magnetic Resonance (NMR) Outputs for Clastic Rocks Using Machine Learning Methods, Examples from North West Shelf and Perth Basin, Western Australia.* Perth : Energies, 2022.

Ayyaz Mustafa, Zeeshan Tariq, Mohamed Mahmoud & Abdulazeez Abdulraheem. 2023. *Machine learning accelerated approach to infer nuclear magnetic resonance porosity for a middle eastern carbonate reservoir .* s.l. : Scientific Reports, 2023.

Srivardhan, . 2022. *Adaptive boosting of random forest algorithm for automatic petrophysical interpretation of well logs.* s.l. : Acta Geodaetica et Geophysica, 2022.

Ahmed M. Elshewey, Mahmoud Y. Shams, Nora El-Rashidy, Abdelghafar M. Elhady, Samaa M. Shohieb, and Zahraa Tarek. 2023. *Bayesian Optimization with Support Vector Machine Model for Parkinson Disease Classification.* s.l. : Sensors, 2023.

Fushiki, . 2011. *Estimation of prediction error by using K-fold cross-validation.* s.l. : Statistics and Computing, 2011.

Damjan Krstajic, Ljubomir J Buturovic, David E Leahy & Simon Thomas. 2014. *Cross-validation pitfalls when selecting and assessing regression and classification models.* s.l. : Journal of Cheminformatics, 2014.

Yongli Zhang, Yuhong Yang. 2015. *Cross-validation for selecting a model selection procedure.* s.l. : Journal of Econometrics, 2015.

Jidesh, Amala Mary VincentP. 2023. *An improved hyperparameter optimization framework for AutoML systems using evolutionary algorithms.* s.l. : Scientific Reports, 2023.

Kazi Ekramul HoqueKazi, Hamoud Aljamaan. 2021. *Impact of Hyperparameter Tuning on Machine Learning Models in Stock Price Forecasting.* s.l. : IEEE, 2021.

Jun Yoneda, Kiyofumi Suzuki, Yusuke Jin, Satoshi Ohtsuki, Timothy S. Collett, Ray Boswell, Yuki Maehara, and Norihiro Okinaka. 2022. *Permeability Measurement and Prediction with Nuclear Magnetic Resonance Analysis of Gas Hydrate-Bearing Sediments Recovered from Alaska North Slope 2018 Hydrate-01 Stratigraphic Test Well.* s.l. : American Chemical Society, 2022.

Razieh Solatpour, Apostolos Kantzas. 2018. *Application of nuclear magnetic resonance permeability models in tight reservoirs.* s.l. : Canadian Society of Chemical Engineers, 2018.

Na Zhang, Fangfang Zhao, Pingye Guo , Jiabin Li, Weili Gong, Zhibiao Guo, and Xiaoming Sun. 2018. *Nanoscale Pore Structure Characterization and Permeability of Mudrocks and Fine-Grained Sandstones in Coal Reservoirs by Scanning Electron Microscopy, Mercury Intrusion Porosimetry, and Low-Field Nuclear Magnetic Resonance.* s.l. : Geofluids, 2018.

Zhen Lu, Aimin Sha, Junfeng Gao, Meng Jia, and Wentong Wang. 2020. *An Improved Predictive Model for Determining the Permeability Coefficient of Artificial Clayey Soil Based on Double T2 Cut-Offs.* s.l. : Earth Materials and Environmental Applications 2020, 2020.

Xing Lei, Liu Xueqin, Liu Huaishan, Qin Zhiliang, and Ma Benjun. 2021. *Research on the Construction of a Petrophysical Model of a Heterogeneous Reservoir in the Hydrate Test Area in the Shenhu Area of the South China Sea (SCS).* s.l. : Geofluids, 2021.