UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

APPLICATIONS OF MACHINE LEARNING METHODS IN THE GENERATION OF SUBSURFACE MEASUREMENTS

A THESIS

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

MASTER OF SCIENCE

By

JIABO HE

Norman, Oklahoma
2018

APPLICATIONS OF MACHINE LEARNING METHODS IN THE GENERATION OF SUBSURFACE MEASUREMENTS

A THESIS APPROVED FOR THE
MEWBOURNE SCHOOL OF PETROLEUM AND GEOLOGICAL ENGINEERING

BY

Dr. Siddharth Misra, Chair

Dr. Deepak Devegowda

Dr. Xingru Wu

Dedicate to my dear mom and dad who gave me countless love and support.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my supervisor, Dr. Siddharth Misra, for his supervision and financial support throughout the duration of my graduate study. It is Dr. Siddharth Misra who helped me in my entire graduate research. I have been learning a lot from him on how to find problems, solve problems, think innovatively and pay attention to details. Without his support and guidance, I will not have the chance to start and complete my graduate research.

Secondly, I would like to appreciate my committee members, Dr. Deepak Devegowda and Dr. Xingru Wu for their help, guidance and suggestions in my research.

Thirdly, I owe a big thank you to all my friends for their help in my research and my life at University of Oklahoma. I would like to say thank you to Hao Li, Yifu Han, Sangcheol Yoon, Pratiksha Tathed and Shiv Prakash Ojha for their help in my research. I would also like to say thank you to Kang Kang, Kailei Liu, Da Zheng, Bin Yuan, Ziyi Xu, Kai Wang, Yao Wang and other friends for their help in my life. I really appreciate the friendship with all of you and I will miss those meaningful and interesting days.

Finally, I am truly grateful to my parents, Weidong He and Beihong Liu, and my whole family. They have been supporting me unselfishly for more than twenty years in every aspect of my study and my life. My father’s insistence and my mother’s optimism are my largest fortunes and best examples in my entire life. I love you!

Table of Contents

List of Tables

List of Figures

Abstract

Machine learning methods have been used in the Oil and Gas industry for about thirty years. Applications range from interpretations of geophysical, well and seismic responses, identification of minerals, analysis of rock samples and cores, fluid properties characterization, formation damage control, risk analysis, to well control (Alegre, 1991). In my thesis, I apply various machine learning methods for generating three well logs in shale formations, namely Nuclear Magnetic Resonance (NMR) T_2 log, Dielectric Dispersion (DD) logs, and sonic travel time logs.

NMR log acquired in geological formations contains information related to fluid-filled pore volume, fluid phase distribution, and fluid mobility. Raw NMR responses of the formation are inverted to generate the NMR T_2 distribution responses in the geological formation, which is further processed to compute the effective porosity, permeability, bound fluid volume, and irreducible saturation of the formation under investigation. I developed two neural-network models that process conventional, easy-to-acquire logs to generate the in-situ NMR T_2 distribution along 300-feet depth interval of a shale reservoir in Bakken Petroleum System (BPS). Following that, we generated DD logs. DD logs acquired in subsurface geological formations generally comprise conductivity (σ) and relative permittivity (ε_r) measurements at 4 discrete frequencies in the range of 10 MHz to 1 GHz. Acquisition of DD logs in subsurface formation is operationally challenging and requires hard-to-deploy infrastructure. I developed three supervised neural-network-based predictive methods to process conventional, easy-to-acquire subsurface logs for generating the 8 DD logs acquired at 4 frequencies. These predictive methods will improve reservoir characterization in the absence of DD logging tool. The predictive methods are tested in three wells intersecting organic-rich shale formations of Permian Basin (PB) and Bakken Shale (BS). Finally, we generated compressional and shear travel time logs (DTC and DTS, respectively) acquired using sonic logging tools. DTC and DTS logs are used to estimate connected porosity, bulk modulus, shear modulus, Young's modulus, Poisson's ratio, brittleness coefficient, and Biot's constant for purposes of geomechanical characterization. Six shallow learning models, namely Ordinary Least Squares (OLS), Partial Least Squares (PLS), Least Absolute Shrinkage and Selection Operator (LASSO), ElasticNet, Multivariate Adaptive Regression Splines (MARS) and Artificial Neural Network (ANN) models, suitable for function approximation problems, are trained and tested to predict DTC and DTS logs. 8481 observations along 4240-feet depth interval of a shale reservoir in Permian Basin (PB) are available for the proposed data-driven application. ANN model performs the best among the six models. Generation of NMR T2 is the computationally most challenging and we had the least amount for data from 220-feet depth interval that made the task even more challenging; nonetheless, we obtained prediction performance of 0.85 in terms of R^2. On the other hand, the generation of dielectric permittivity and conductivity dispersion logs was slightly lower in terms of computational cost as compared to NMR T2 generation, we had data from 2200-feet depth interval, and prediction performance for this log generation task was 0.79 in terms of R^2 in average. Generation of DTC and DTS logs is computationally easiest among the three tasks, we had data from 4240-feet depth interval, and the prediction performance was 0.86 in terms of R^2 in average.

Chapter 1: Introduction

This thesis presents work performed for a Master of Science degree that was conducted at Mewbourne School of Petroleum and Geological Engineering of the University of Oklahoma. The research presented the application of machine learning methods in generation of three well logs, namely Nuclear Magnetic Resonance (NMR) T_2 log, Dielectric Dispersion (DD) logs, and sonic travel time logs. The research work was done in close collaboration and supervision of Dr. Siddharth Misra.

Machine learning is a subfield of artificial intelligence exploring the study of algorithms that can learn from and make predictions on data or images (Russell and Norvig, 2016). In oil and gas industry, well logs provide information of formations in the subsurface, including data and images (Asquith et al., 2004). In this thesis, different models and methods are developed and compared for the various cases and sensitivity analyses are done on these models and methods. The following sections present the motivation behind the study as well as the objectives, background, approach and outline of the thesis.

1.1 Motivation and Problem Statement

Oil and gas engineers estimate subsurface properties using well logs. Sometimes, it can be financially and operationally challenging to deploy certain logging tools, such as nuclear magnetic resonance (NMR), dielectric dispersion (DD) and sonic logging tools. More importantly, there does not exist physical or data-driven models that can process/combine the easy-to-acquire logs, which individually have lower information content, to obtain logs and responses that have higher information content and closer to human perception. In this study we will process easy to acquire logs to synthesize hard-to-acquire logs, such as dielectric dispersion, sonic, and NMR logs.

NMR log contains information related to fluid-filled pore volume, fluid phase distribution, and fluid mobility (Salazar and Romero, 2001). NMR T_2 distribution is generated from raw NMR responses in the geological formation, which approximates the in-situ pore size distribution. DD log measures conductivity and permittivity dispersion at different frequencies, which provides water-filled porosity and formation water salinity (Hizem et al., 2008). DTC and DTS logs acquired using sonic logging tools are used to estimate connected porosity, bulk modulus, shear modulus, Young's modulus, Poisson's ratio, brittleness coefficient, and Biot's constant for purposes of geomechanical characterization (Maleki et al., 2014). As a result, all of them are crucial to oil and gas industry although it is difficult to obtain them in some wells. Considering the importance of these logs, one possible method is to predict them with machine learning methods using other conventional, easy-to-acquire logs. People can obtain them without deploying these logging tools after building proper and efficient machine learning models.

1.2 Objective

(1) Two different ANN models are built to predict NMR T_2 distribution responses with conventional, easy-to-acquire logs in Bakken Petroleum System (BPS). Prediction performance and computational cost of two models are compared. Three set of data are used to verify the ANN models.

(2) Three predictive methods are proposed to generate conductivity and permittivity dispersion logs in three wells in Permian Basin (PB) and Bakken Shale (BS). Prediction performance and computational cost of three methods in different cases

are compared. Sensitivity analysis of inputs, outputs and noise to the predictive methods is done.

(3) Six shallow learning models are compared to predict compressional and shear travel time logs (DTC and DTS, respectively) in PB. All models are trained and tested in Well 1 and deployed in Well 2. Prediction performance and computational cost of six models are compared. Sensitivity analysis of noise and dataset to the ANN model is done.

1.3 Outline of Thesis

Chapter 2 is literature review, which provides the former research of machine learning methods applied in generation of well logs.

Chapter 3 predicts NMR T_2 distribution responses by two different ANN models with conventional, easy-to-acquire logs in BPS. Flags are assigned using the KNN method to add more inputs for two models. Two models are compared in both predictive performance and computational time. The first predictive model exhibits better performance with higher computational cost. Three NMR-related sets of data (ϕ_N, $T_{2,gm}$ and $T_{2,gm}\phi_N^2$) derived from the generated fluid-filled pore size distribution establish the robust performance of the predictive models.

Chapter 4 predicts DD log by three different predictive methods with conventional, easy-to-acquire logs in three wells in PB and BS. the prediction performances of the second predictive method are worse for conductivity dispersion logs and better for permittivity dispersion logs, respectively, as compared to those of the first predictive method. The third method has best prediction performance for both conductivity and permittivity dispersion logs. All three methods perform well in Well 1 and 2 but perform badly in Well 3 because of fewer samples and higher water salinity in Well 3.

Chapter 5 trains and tests six shallow learning models (OLS, PLS, LASSO, ElasticNet, MARS and ANN models) to predict DTC and DTS logs in PB. ANN model performs the best among the six models. The trained ANN model is deployed in another well drilled in the same reservoir with comparable good prediction performance. The complex structure of the ANN model with its multiple neurons and layers make it the best performing model among the six models.

Chapter 6 is the conclusion of this thesis, including the prediction performance of all models in different cases, advantages and disadvantages of every model and sensitivity analysis results of all cases with respect to the research area.

Chapter 2: Background and Literature Review

2.1 Introduction to the Well Logs

2.1.1 Nuclear Magnetic Resonance (NMR) Log

NMR T_2 distribution response of a geological formation is captured during the acquisition of NMR log. In hydrocarbon reservoirs, T_2 is the transverse relaxation time of hydrogen nuclei of the fluid phases that fill the pores. NMR T_2 distribution response is governed by the fluid-filled pore volume, fluid phase distribution, and fluid mobility. NMR T_2 distribution approximates the pore size distribution. Concentration of T_2 distribution around small T_2 times originate due to small-sized pores. Operational challenges during NMR tool deployment impedes its extensive application and NMR log acquisition. The first objective of this thesis is to synthetically generate NMR T_2 distribution from conventional and easy-to-acquire logs, so that NMR T_2 distribution can be obtained in wells where NMR log is not available due to well conditions or economics, such as the lateral well section of a well in a shale reservoir or small-diameter boreholes in deep HPHT reservoirs.

2.1.2 Dielectric Dispersion Logs

Dielectric dispersion (DD) response of a subsurface geological formation is acquired using the wireline DD logging tool that is run in an open-hole well intersecting the formation (Han et al., 2017). Conductivity measures the ability of rocks and fluids to conduct an electric current while permittivity is determined by the ability of rocks and fluids to polarize in response to the electric field. DD logs measure conductivity and permittivity dispersion at different frequencies, usually in the range of 10 MHz to 1 GHz.

Conductivity dispersion and permittivity dispersion can provide water-filled porosity, formation water salinity and water saturation.

2.1.3 Sonic Travel Time Logs

Sonic logging tools transmit sonic waves that propagate from multiple sources to receivers. Compressional and shear travel time logs (DTC and DTS, respectively) can be estimated from the waveforms recorded at the receiver, and both the logs have high degree of correlation (Willis and Toksoez, 1983). Compressional waves are longitudinal waves while shear waves are transverse waves. They travel at different speed in solid, liquid and gas so that different lithology, different porosity and different combinations of liquid and gas in pores result in different DTC and DTS logging data. Higher speed of compressional waves than shear waves lead to smaller DTC values than DTS values. DTC and DTS values are affected by porosity the most and they tend to be large under the circumstances of large porosity and tend to be small when porosity is small. As a result, DTC and DTS can be used to estimate different reservoir properties for geomechanical characterization, such as porosity, various moduli (bulk modulus, shear modulus and Young's modulus), Poisson's ratio and so on.

2.2 Previously Developed Data-Driven Models for Subsurface Applications

2.2.1 Nuclear Magnetic Resonance (NMR) Log

ANN have been applied to improved log-based shale gas and coal reservoir characterization. Bhatt and Helle (2002) built committee neural networks for porosity and permeability predictions by processing well logs from the North Sea, where porosity-prediction ANN uses sonic, density and resistivity logs as inputs and permeability-prediction ANN uses density, gamma ray, neutron porosity and sonic logs as inputs. Al-

Bulushi et al. (2007) predicted water saturation in the Haradh sandstone formation using ANNs with wireline logs as inputs, which included density, neutron, resistivity and photoelectric logs. Recently, Mahmoud et al. (2017) built an ANN model to predict TOC for Barnett shale based on resistivity, gamma ray, sonic transit time and bulk density logs. The model was then applied to estimate TOC for Devonian shale. Several studies applied ANN algorithms in predicting NMR-T_2-derived parameters. Salazar and Romero (2001) predicted NMR porosity and permeability using ANNs with gamma ray, resistivity and neutron logs as inputs in a carbonate reservoir. Mohaghegh et al. (2001) synthesized magnetic resonance logs such as free fluid, irreducible water and effective porosity using ANNs with SP, gamma ray, caliper and resistivity as inputs. Later, Elshafei and Hamada (2009) predicted permeability using Bulk Gas model and ANN model separately and validated the results against permeability measurements on core samples. Labani et al. (2010) estimated free-fluid-filled porosity and permeability using a committee machine with intelligent systems (CMIS) in the South Pars gas field. CMIS combines the results of Fuzzy Logic, Neuro-Fuzzy and Neural Network algorithms for overall estimation of NMR log parameters from conventional log data. Recently, Golsanami et al. (2014) predicted eight bin porosities and T_2 Logarithmic Mean ($T_{2,LM}$) values of NMR T_2 distribution using intelligent models for characterization in the Asmari formation. I have not come across published research work on developing and applying ANN-based predictive models of the generating the entire NMR T_2 spectral response, that approximates the fluid-filled pore size distribution in hydrocarbon-bearing reservoir. The prediction of NMR T_2 distribution help people estimate porosity, permeability, lithology and minerals, quantity of hydrocarbons and so on. The success of predicting NMR T_2

distribution will avoid of predicting those formation properties one by one with different models. Those properties can be estimated by predicted NMR T_2 distribution.

In relation to the second predictive model implemented in my thesis, Genty et al. (2007) fitted NMR T_2 distribution response acquired in a carbonate reservoir with multiple Gaussian (or normal) distributions that can be characterized with three parameters (α, μ, σ) for each distribution. In their case, T_2 distributions required three Gaussian components that needed 9 parameters for purposes of fitting. Genty et al. (2007) utilized these fitted parameters to identify genetic pore types in the carbonate reservoir. Di (2015) identified different lithofacies in a tight oil reservoir based on the parameters estimated from fitting the T_2 distribution similar to that proposed by Genty et al. (2007). I use a similar approach to generate 6 parameters for training and testing the models for NMR T_2 prediction.

2.2.2 Dielectric Dispersion Logs

Hizem et al. (2008) was first to report the acquisition of borehole-based DD logs with a new-generation dielectric tool for the continuous measurement of DD logs at 1-in vertical resolution at four discrete frequencies. Pad-based transmitters on the dielectric tool remain in contact with the geological formation. The transmitters send electromagnetic (EM) waves of known magnitude and phase in the range of 10 MHz to 1 GHz into the formation. The EM waves travel through the formation, reach the pad-based receivers, and the attenuation and phase shift of the waves due to the material properties of the formation are recorded. Following that, the wave attenuation and phase shift are inverted using the tool-physics forward model to compute multifrequency conductivity (σ) and relative permittivity (ε_r) dispersions of the formation in the frequency range of 10 MHz to 1 GHz. Finally, a valid multifrequency geo-electromagnetic mixing model or mechanistic polarization-dispersion model is applied on the measured conductivity and permittivity dispersion logs to estimate water saturation, bound water saturation, salinity, clay-exchange capacity, and textural parameters (Tathed et al., 2018; Han and Misra, 2018).

Similar, synthetic NMR T2 log generation under constraints was reported by Li and Misra (2017a) using variational autoencoder based neural networks (Li and Misra, 2017b). It should be noted that for generating synthetic DD logs in deviated well, the predictive methods need to be trained on logs form deviated wells along with transfer learning from methods trained on logs from vertical wells.

Brovko et al. (2009) proposed an improvement in microwave imaging technique by ANN-assisted reconstruction of 2-D complex permittivity profiles in dielectric samples placed in a waveguide system. The spatial distributions of the dielectric constant were estimated as continuous functions whose parameters were predicted using ANN models. Chen et al. (2011) developed a back propagation (BP) ANN model to compute the effective complex permittivity of liquid materials (organic solvents) by processing the measured scattering parameters, such that the relative errors in prediction were less than 5%. Hasan et al. (2011) developed two ANN models to separately predict the real and imaginary permittivity components of thermos-responsive materials using the magnitude and phase of reflection coefficients for different frequencies from 2.5 GHz to 5 GHz. A similar dielectric prediction work used Complex-Valued Neural Network (CVNN) models. CVNNs were designed specially to predict complex-valued parameters. CVNN was first proposed by Aizenberg et al. in former Soviet Union in 1971 (Nitta, 2009). Yang et al. (2005) applied CVNN models to landmine detection and classification applications using data predicted by Ground Penetrating Radar. With phase information of scattering parameters included as inputs for CVNN, better prediction performance was obtained in comparison to real-valued Neural Network models.

2.2.3 Sonic Travel Time Logs

Many researchers predicted DTS (or V_s reciprocal of DTS) with DTC (or V_p, reciprocal of DTC) and other conventional logs as inputs to couple robust relations between DTS and DTC. Empirical equations were used to estimate DTS from lithology, porosity, S_w and DTC, verified by two sets of logging data in the offshore Gulf of Mexico clastics with accuracy in terms of R2 equal to 0.81 and 0.76 respectively (Greenberg and Castagna, 1992).. Intelligent systems including fuzzy logic (FL), neuro-fuzzy (NF) and Artificial Neural Network (ANN) models predicted DTS successfully with NPOR, DTC, GR, RHOZ, and deep laterolog resistivity (Rlld) as inputs in the sandstone reservoir of Carnarvon Basin in Australia (Rezaee et al., 2007). Measured error using mean squared error (MSE) was about 0.05 for three models respectively, which were close to each other. Empirical correlations and machine learning methods were compared to predict DTS in the carbonate reservoir in Iran (Maleki et al., 2014). Empirical correlations used DTC to calculate DTS in empirical equations, including Castagna equation, Brocher equation and Carroll equation, with accuracy in terms of R2 equal to 0.92, 0.86 and 0.88, respectively. Machine learning methods, including SVM and ANN, predicted DTS with DTC, RHOZ, GR, effective porosity (PHIT), true formation resistivity (Rt) and DCAL as inputs, with accuracy in terms of R2 equal to 0.94 and 0.88 respectively. In addition, some people also predicted DTC and DTS together. Empirical equations were implemented to derive DTC and DTS with GR, lithology and porosity in Oregon basin in Wyoming with high accuracy (Iverson and Walker, 1992). Xu-White model was proposed to estimate V_p and V_s of shaly sandstones from porosity and shale content with high accuracy (Xu and White, 1995). A key step in Xu-White model was to estimate dry rock bulk and shear moduli for the sand/shale mixture, which was improved in computational time by assuming constant Poisson's ratio (Keys and Xu, 2002). Numerical experiments showed a close match, less than 2% in terms of relative error (RE), between velocities obtained with the approximations and those computed with the original differential effective medium method in Xu-White model. In another work, the Thomas-Stieber approach to petrophysical analysis of thin beds and the Dvorkin and Gutierrez Sand/Shale rock physics models were applied to predict V_p and V_s in thin beds using mineral properties, rock porosity and Vsh for initial parameters, which had potential but considerable limitations (Baines et al., 2008). Committee machine with intelligent systems (CMIS) were used to predict V_p V_s and stoneley wave velocity (V_{st}) at the same time in the Asmari formation, the carbonate reservoir of Iranian oil field with NPOR, RHOZ, Rt and Vsh as inputs (Asoodeh and Bagheripour, 2012). FL, NF and ANN were included in CMIS and the system predicted V_p, V_s and V_{st} with accuracy in terms of R2 equal to 0.93, 0.89 and 0.74, respectively. However, no research has been done to balance the simplicity of models and prediction performance for sonic log so far.

2.3 Conclusions

There are several subsurface environments, operational challenges, and project economics scenarios where the three logs cannot be acquired, for example in the lateral and deviated sections of a well in shale reservoir, high-pressure high-temperature wells in deep reservoirs, and irregularly sized boreholes in carbonate reservoirs. For those situations, machine learning methods are applied to generate these logs by processing conventional, easy-to-acquire logs in the absence of target logs.

In the first research, the first predictive model performs more accurately, exhibiting median R^2 of 0.8549 during testing, compared to the second one, exhibiting median R^2 of 0.7584. However, the second model has lower computational cost compared to the first model. In the second research, prediction performance of the second method is 8.5% worse for conductivity dispersion and 6.2% better for permittivity dispersion than the first one. Prediction performance of the third method is 0.8% better for conductivity dispersion and 8.5% better for permittivity dispersion than the first one. Training the third predictive method in one well and then deploying it in another well for generating the 8 DD logs is feasible, such that the NRMSE of conductivity dispersion logs drops by 6% and that of permittivity dispersion logs drops by 4.5% compared to the baseline. In the third research, six shallowing learning models are selected to predict DTC and DTS at the same time in PB. OLS, PLS, LASSO and ElasticNet models are four linear regression models while MARS and ANN can also deal with problems of high non-linearities. After comparison, ANN model performs the best both in Well 1 and Well 2.

Chapter 3: Generation of In-Situ NMR T_2 Distribution in Bakken Petroleum System Using Neural Networks

3.1 Theory and Methodology

3.1.1 Introduction of the Bakken Petroleum System (BPS)

BPS is a hybrid play composed of both conventional and unconventional elements (Simpson et al., 2015), a large portion of which is in northeastern Montana and northwestern North Dakota (Figure 3.1). The well under investigation in this thesis is drilled in Hawkeye field, North Dakota. The formations intersected by the well under investigation are somewhat similar to the distribution shown in Figure 3.1. Conventional plays in this well consist of separate reservoir intervals (Middle Bakken shale, Three Forks formation), source rock intervals (Lower Bakken and Upper Bakken shales), and other intervals. Upper Bakken Shale (UBS), Middle Bakken Shale (MBS), Lower Bakken Shale (LBS) and Three Forks (TF) formation are consecutive intervals in BPS from top to bottom, which are my target intervals (Figure 3.2). Oil and gas produced in the UBS and LBS got accumulated in the MBS and TF formations. These intervals are distinctly different and display highly heterogeneous distributions of reservoir properties. 7 intervals intersected by the science well are targeted in my research with distinct lithologies, which include black shale, siltstone, sandstone, dolostone and dolo-mudstone. Different minerals as present in these intervals, such as quartz, K-feldspar, plagioclase feldspars, illite, dolomite, calcite, kaolinite, and pyrite.

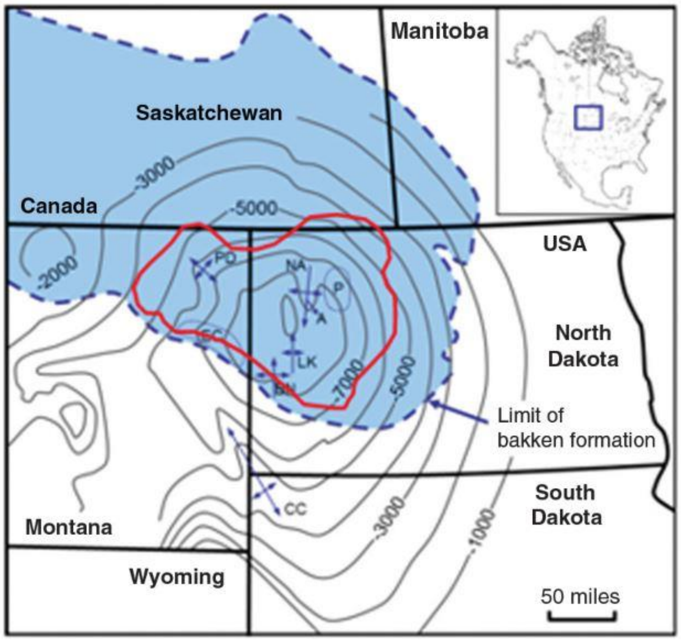


Figure 3.1 The Bakken formation in the Williston basin, underlying sections of North Dakota, northern Montana, southern Saskatchewan and southwestern Manitoba (Pirie et al., 2016)

The UBS and LBS formations are black shale formations with deposit time spanning from the Late Devonian period to Early Mississippian period. Similar depositional conditions produce two similar deposits although they belong to different depositional episodes. Siliceous portions, calcareous portions and pyrite are found in both these formations. The mineralogical composition of the two formations, based on X-ray Diffraction (XRD) analyses, is dominated by quartz, K-feldspar and plagioclase feldspars. The clay mineral content is from 13 to 43 vol% dominated by illite; the carbonate content is from 2 to 13 vol% dominated by dolomite; and TOC content is from 10 to 20 wt% (Simpson et al., 2015). The MBS formation displays a range of grain size sorting from poorly sorted, argillaceous siltstone to moderately well sorted fine-grained sandstone (Simpson et al., 2015) and is more complex than UBS and LBS formations. Different grain size, depositional textures and diagenetic calcite cement lead to more detailed division of MBS, which are Upper Middle Bakken (UMB) and Lower Middle Bakken (LMB). UMB contains better reservoir quality with well sorted, fine-grained sandstone, whereas LMB contains more bioturbated, silt-dominated, shallow-marine deposits. The TF formation is a dolostone with alternating porous dolo-siltite facies. The interlaminated TF dolostone is interbedded with clay-rich, conglomeratic dolo-mudstone, which marks stratigraphic intervals that partition TF formation into four distinct sequences from top to bottom. TF1 is the principle oil-bearing interval of TF. TF2 is also oil-bearing but only locally charged with oil, mainly in the center of the basin. It is rare to find oil in TF3 and the remaining TF4 is non-reservoir dolo-mudstone (Simpson et al., 2015).

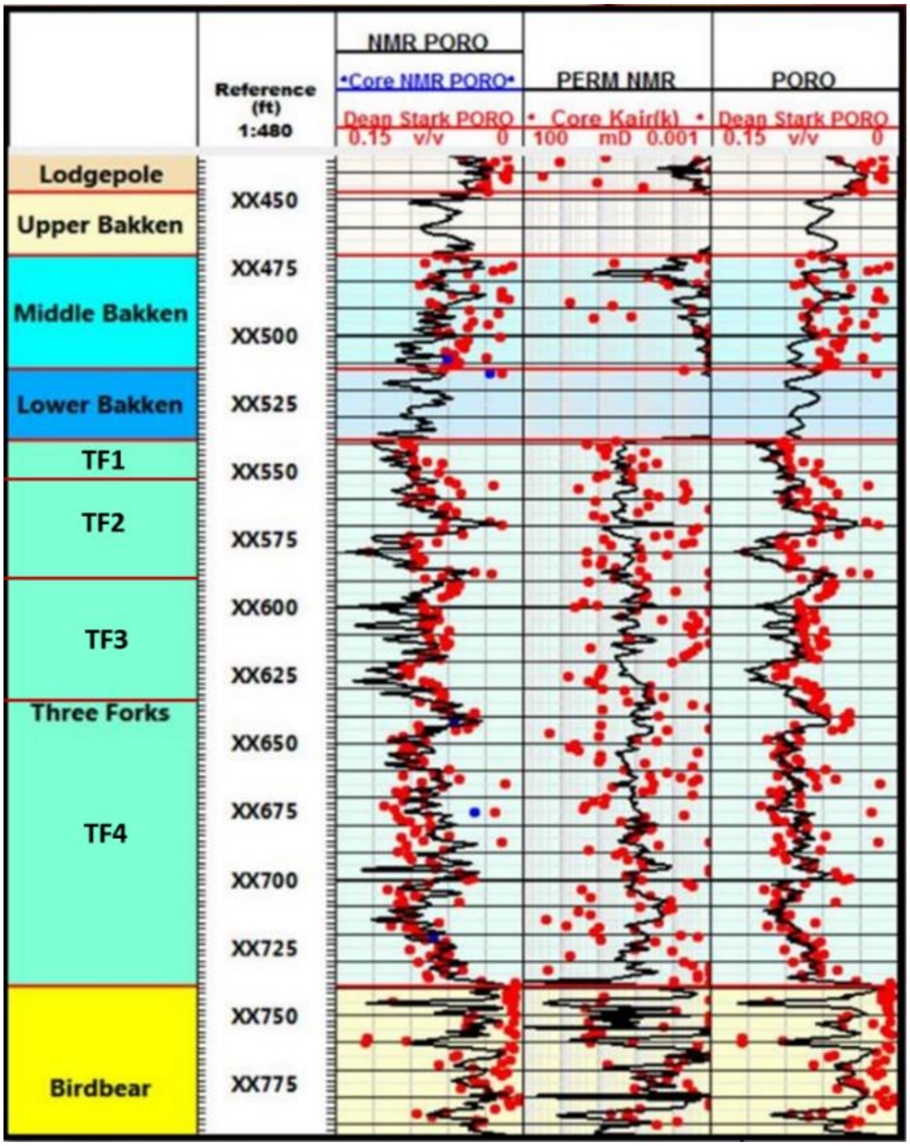


Figure 3.2 Seven distinct intervals in BPS, consisting of the UBS, MBS, LBS, TF1, TF2, TF3 and TF4 formations, as adapted from Simpson et al. (2015)

3.1.2 Data Preparation and their Statistical Description

Data preparation and data preprocessing were done prior to the application of the predictive models (Figure 3.3). The first ANN model predicts T_2 distribution discretized into 64 bin amplitudes, whereas the second ANN model predicts the six parameters that characterize the T_2 distribution as a sum of two Gaussian distributions. The second model requires the generation of the six characteristic parameters prior to prediction. Evaluation of the accuracies of the two predictive models is the crucial last step to ensure robustness and reliability of the models (Figure 3.3).

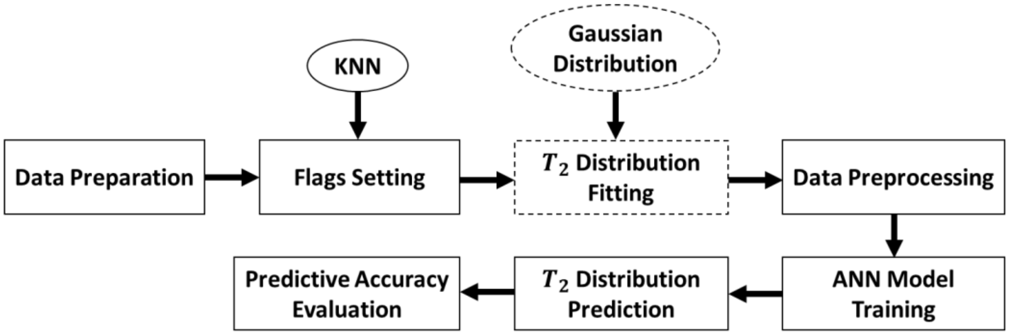


Figure 3.3 Flowchart for the predictive models, where the actions represented inside the dashed boundary are involved in the second predictive model

Target depth interval was selected from XX676 ft to XX985 ft comprising 7 intervals of a shale reservoir system. Overall, inputs to the two predictive models include 12 conventional logs, 10 inversion-derived logs, and 5 qualitative flags. The conventional logs include gamma ray (GR) log sensitive to volumetric shale concentration, induction resistivity logs measured at 10-inch (AT10) and 90-inch (AT90) depths of investigation sensitive to the volumes of connate hydrocarbon and brine, and neutron (NPOR) and density porosity (DPHZ) logs that are influenced by the formation porosity. Other conventional logs include photoelectric factor (PEFZ) log indicating the formation interval, VCL log measuring the volume of clay, and RHOZ log sensitive to the formation density. Finally, sonic logs, including Delta-T Shear (DTS), Delta-T Compressional (DTC) and Shear to Compressional Velocity Ratio (VPVS), that sense the geomechanical properties of the interval and rock textures are also used as inputs. In addition, 10 input logs were generated using a commercial inversion software that processed the 12 conventional logs. Six of the ten inversion-derived logs are presented in Tracks 8 and 9 in Figure 3.4, which are mineral content (of quartz, calcite, and dolomite) and fluid saturation logs. NMR T_2 distributions (Track 10) are also used to train and test ANN models and to generate the 6 parameters characterizing the T_2 distributions. Environmental correlations and depth corrections were performed on all these logs prior to processing.

Table 3.1, 3.2 and 3.3 list statistical descriptions of all the input logs. Median GR of UBS and LBS are much higher than those of other intervals, which indicates dominance of organic shale in UBS and LBS. AT10 and AT90 are much higher in UBS and LBS than those in other intervals, which indicates high hydrocarbon in these two intervals. DPHZ and NPOR are also the largest in UBS and LBS for the same reason. Statistical parameters for most of the input logs in UBS and LBS are similar, whereas those in TF1, TF2, TF3, and TF4 are similar. Statistical parameters for most of the input logs in MBS are closer to those of TF formation rather than those in UBS and LBS formations. Coefficient of variation is the value of standard deviation over mean value (S_d/μ), which is used to measure the dispersion of data. NaN values in Table 3.2 represent mean values of 0 when calculating coefficient of variation. Most input logs in the 7 intervals have reasonable coefficient of variation. Skewness is the measurement of asymmetry of data about its mean. NaN values in Table 3.3 represent all same values when calculating skewness. A few inputs have both large coefficient of variation and large skewnesses in all intervals such as AT10 and AT90, which indicates both high dispersion and high asymmetry of them.

Table 3.1 Median value of every input in every interval.

	Median						
	UBS	**MBS**	**LBS**	**TF1**	**TF2**	**TF3**	**TF4**

GR	**446.800**	**78.520**	**775.400**	**89.570**	**95.260**	**80.730**	**90.645**
VPVS	**1.625**	**1.675**	**1.659**	**1.716**	**1.768**	**1.774**	**1.776**
DCAL	**-0.275**	**-0.243**	**-0.275**	**-0.244**	**-0.248**	**-0.259**	**-0.260**
AT10	**411.100**	**8.887**	**137.500**	**3.979**	**2.555**	**2.268**	**2.236**
AT90	**169.500**	**8.394**	**74.280**	**3.582**	**2.214**	**2.167**	**2.025**
DTC	**96.300**	**63.405**	**99.080**	**64.890**	**63.755**	**63.040**	**64.245**
DTS	**155.300**	**107**	**163.700**	**113.300**	**112.100**	**111.900**	**114.700**
DPHZ	**0.236**	**0.054**	**0.248**	**0.033**	**0.020**	**0.004**	**0.021**
NPOR	**0.275**	**0.073**	**0.284**	**0.130**	**0.142**	**0.129**	**0.141**
PEFZ	**2.967**	**3.253**	**2.850**	**3.052**	**3.125**	**3.572**	**3.703**
RHOZ	**2.307**	**2.618**	**2.286**	**2.653**	**2.677**	**2.703**	**2.675**
VCL	**1**	**0.479**	**1**	**0.581**	**0.583**	**0.478**	**0.575**
Illite	**0.294**	**0.003**	**0.313**	**0.071**	**0.074**	**0.278**	**0.297**
Chlorite	**0.002**	**0.020**	**0.010**	**0.01**	**0.046**	**0**	**0**
Bound water	**0.032**	**0.004**	**0.033**	**0.010**	**0.015**	**0.046**	**0.049**
Quartz	**0.298**	**0.340**	**0.233**	**0.144**	**0.129**	**0.079**	**0.065**
K-feldspar	**0.015**	**0.173**	**0.058**	**0.200**	**0.225**	**0.071**	**0.121**
Calcite	**0**	**0.246**	**0**	**0.030**	**0.058**	**0.078**	**0.110**
Dolomite	**0**	**0.132**	**0**	**0.419**	**0.391**	**0.285**	**0.258**
Anhydrite	**0.037**	**0**	**0.069**	**0**	**0**	**0.096**	**0.031**

Unflushed water	**0.00001**	**0.029**	**0.001**	**0.035**	**0.049**	**0.005**	**0.004**
Unflushed oil	**0.007**	**0.020**	**0.005**	**0.017**	**0**	**0**	**0**

Table 3.2 Coefficient of variation of each input log in each interval.

	S_d/μ						
	UBS	**MBS**	**LBS**	**TF1**	**TF2**	**TF3**	**TF4**
GR	**0.287**	**0.219**	**0.179**	**0.260**	**0.158**	**0.203**	**0.195**
VPVS	**0.024**	**0.037**	**0.023**	**0.022**	**0.019**	**0.025**	**0.028**
DCAL	**-0.084**	**-0.078**	**-0.060**	**-0.073**	**-0.066**	**-0.060**	**-0.048**
AT10	**1.028**	**0.618**	**0.883**	**1.468**	**0.290**	**0.339**	**4.123**
AT90	**0.684**	**0.323**	**0.697**	**0.885**	**0.319**	**0.595**	**3.611**
DTC	**0.104**	**0.069**	**0.067**	**0.087**	**0.034**	**0.044**	**0.071**
DTS	**0.097**	**0.046**	**0.074**	**0.079**	**0.035**	**0.027**	**0.050**
DPHZ	**0.329**	**0.269**	**0.179**	**0.531**	**0.480**	**-16.655**	**3.376**
NPOR	**0.321**	**0.449**	**0.143**	**0.337**	**0.138**	**0.315**	**0.307**
PEFZ	**0.142**	**0.116**	**0.076**	**0.026**	**0.087**	**0.101**	**0.073**
RHOZ	**0.051**	**0.009**	**0.031**	**0.010**	**0.007**	**0.018**	**0.022**
VCL	**0.102**	**0.319**	**0**	**0.334**	**0.170**	**0.240**	**0.261**
Illite	**0.399**	**1.461**	**0.299**	**0.814**	**0.856**	**0.228**	**0.320**

Chlorite	**1.381**	**0.782**	**1.171**	**1.079**	**0.659**	**2.673**	**1.836**
Bound water	**0.411**	**0.614**	**0.251**	**0.650**	**0.679**	**0.226**	**0.315**
Quartz	**0.233**	**0.229**	**0.220**	**0.331**	**0.314**	**0.398**	**0.579**
K-feldspar	**1.673**	**0.355**	**0.963**	**0.306**	**0.158**	**0.616**	**0.493**
Calcite	**2.691**	**0.510**	**6.072**	**0.965**	**1.004**	**0.728**	**0.517**
Dolomite	**2.048**	**0.468**	**2.333**	**0.189**	**0.339**	**0.270**	**0.285**
Anhydrite	**0.856**	**NaN**	**0.338**	**NaN**	**1.771**	**1.002**	**1.327**
Unflushed water	**2.194**	**0.154**	**2.151**	**0.322**	**0.504**	**0.449**	**1.338**
Unflushed oil	**0.848**	**0.604**	**0.536**	**0.433**	**2.802**	**5.774**	**5.121**

Table 3.3 Skewness of each input log in each interval.

	Skewness						
	UBS	**MBS**	**LBS**	**TF1**	**TF2**	**TF3**	**TF4**
GR	**-1.571**	**-0.835**	**-1.524**	**-0.430**	**0.371**	**-1.262**	**-0.130**
VPVS	**0.549**	**0.787**	**-0.626**	**0.751**	**-1.438**	**2.491**	**1.095**
DCAL	**1.748**	**0.087**	**1.407**	**-0.043**	**0.046**	**-0.056**	**-0.014**
AT10	**1.750**	**2.722**	**2.835**	**3.576**	**1.918**	**1.799**	**8.971**
AT90	**1.503**	**2.162**	**0.473**	**1.636**	**2.175**	**1.830**	**8.198**
DTC	**-2.138**	**0.246**	**-2.814**	**-0.744**	**0.020**	**-1.078**	**-0.723**

DTS	**-1.816**	**1.279**	**-1.499**	**-0.817**	**0.026**	**-0.319**	**-0.795**
DPHZ	**-1.676**	**0.022**	**-3.190**	**-0.484**	**0.312**	**-2.408**	**-1.293**
NPOR	**-1.442**	**-0.758**	**-0.477**	**-0.430**	**0.238**	**-0.771**	**-1.073**
PEFZ	**1.180**	**0.790**	**2.130**	**0.277**	**0.808**	**2.059**	**1.079**
RHOZ	**1.675**	**-0.016**	**3.190**	**0.490**	**-0.308**	**2.409**	**1.295**
VCL	**-3.845**	**-1.113**	**NaN**	**-0.417**	**0.687**	**-0.791**	**-0.383**
Illite	**-1.630**	**1.487**	**-0.567**	**0.367**	**0.658**	**-1.042**	**-0.572**
Chlorite	**1.073**	**-0.017**	**0.888**	**0.727**	**0.385**	**2.870**	**2.026**
Bound water	**-1.458**	**0.799**	**-0.785**	**0.191**	**0.887**	**-1.053**	**-0.617**
Quartz	**0.081**	**-0.341**	**-0.768**	**0.217**	**-0.334**	**-0.322**	**-0.290**
K-feldspar	**2.035**	**-0.754**	**1.257**	**-0.004**	**-0.204**	**-0.070**	**-0.123**
Calcite	**2.778**	**1.007**	**6.373**	**0.433**	**1.224**	**1.005**	**0.013**
Dolomite	**1.972**	**0.598**	**2.857**	**0.247**	**-0.574**	**-0.810**	**0.913**
Anhydrite	**0.495**	**NaN**	**-0.811**	**NaN**	**2.133**	**1.886**	**1.425**
Unflushed water	**2.992**	**-0.006**	**3.852**	**-0.440**	**-0.915**	**0.396**	**3.434**
Unflushed oil	**3.003**	**-0.201**	**-0.054**	**0.604**	**2.675**	**6.137**	**6.364**

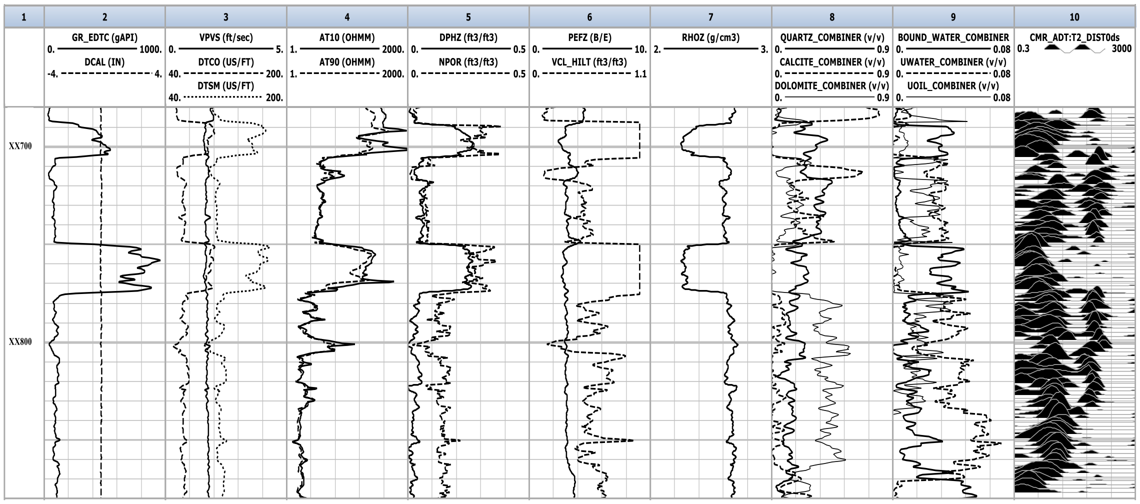
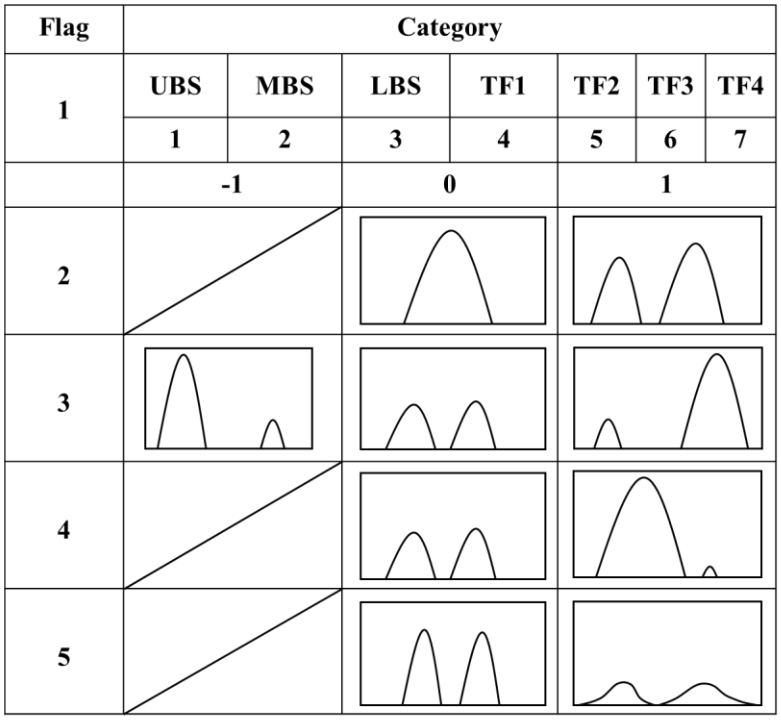


Figure 3.4 Track 1 is depth, Track 2 contains gamma ray and Caliper, Track 3 contains DTS, DTC and VPVS, Track 4 is induction resistivity logs at 10-inch (AT10) and 90-inch (AT90) depth, Track 5 contains density porosity and neutron porosity, Track 6 is formation photoelectric factor and volume of clay, Track 7 is formation density, Tracks 8 & 9 contains 6 inversion-derived logs of mineral content and fluid saturation, and Track 10 is NMR T$_2$ distributions.

3.1.3 Categorization of Depths Using Flags

After selecting the primary logs to be used as inputs for the neural-network based prediction, I assign five qualitative flags to each depth and use them as synthetic discrete-valued logs to improve the prediction accuracy. This can be considered as an effort to categorize depths based on five lithological/geological/textural features. Refer to Figure 3.5 for a qualitative understanding of Flags 1-5. The value of Flag-1 is an integer ranging from 1 to 7 identifying seven distinct intervals based on different lithologies and mineralogical compositions at a given depth. Flag-2 is either 0 or 1 identifying unimodal and bimodal pore size distribution, respectively, at the given depth. Flag-3 is an indicator of pore sizes in a bimodal system, such that its value is -1, 0, or 1 that identify the abundance of small pores, comparable volumes of small and large pores, and abundance of large pores, respectively. Similar to Flag-3, Flag-4 is an indicator of relative abundance of pores of certain pore size in a bimodal system, such that Flag-4 is assigned a value of

1 when certain pore size (either small pores or large pores) is negligible or else it is assigned to be 0. Those assigned as 1 of Flag-4 are bimodal distributions but can be regarded as unimodal distributions. Flag-5 defines the deviation of pore sizes around the two dominant pore sizes of a bimodal distribution, such that 1 indicates that the spreads around the two peaks are wide and 0 indicates either a unimodal distribution or a narrow spread around the two dominant pore sizes. In brief, Flag-1 classifies intervals based on interval and lithology; Flag-2 identifies number of peaks in the pore size distribution; Flag-3 identifies the dominant pore sizes in bimodal pore systems; Flag-4 checks if certain pore sizes can be neglected; and Flag-5 captures difference in the deviation of pore size distributions. These flags help improve the prediction performance as they provide supporting information to the predictive models about specific formation intervals.



Flag	Category						
1	UBS	MBS	LBS	TF1	TF2	TF3	TF4
	1	2	3	4	5	6	7
	-1		0		1		
2							
3							
4							
5							

Figure 3.5 Values assigned to each flag and the corresponding NMR T_2 distribution representations.

These flags are created using k-nearest neighbor (KNN) algorithm. My hypothesis is Flags 1-5 are required to improve the performance of the predictive models. Flags 2-5 can be generated relatively easily for depths where T_2 distribution responses have been acquired. However, for the T_2 distribution prediction work, Flags 2-5 needs to be predicted prior to the primary objective of generating the T_2 distribution. To that end, KNN classification models, specially designed for predicting categories, are used to generate Flags 2-5. The goal of KNN algorithm is to first relate the available easy-to-acquire logs to the NMR T_2 distribution, which can then be easily related to Flags 2-5. Once the KNN algorithm is well trained and tested for accurately relating the easy-to-acquire logs to Flags 2-5 in the presence of NMR T_2 distributions, the KNN model can be subsequently used to generate Flags 2-5 in the absence of NMR T_2 distribution. 22 input logs are all used as inputs to predict Flags 2-5 one by one with the KNN classification models. KNN algorithm classifies new cases based on a similarity measure, and similarity is measured by distances between cases. Based on training data, nearest neighbors for all testing data can be found by KNN algorithm, which makes it plausible to predict outputs of testing data using these neighbors in KNN classification models. I use Hamming Distance Function (D_H) to find nearest neighbors that is expressed as

$$D_H = \sum_{i=1}^{k} |x_i - y_i| \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots......(3.1)$$

where x and y are two random cases. k value is selected as 3 among possible values ranging from 1 to 10; in other words, Flags 2-5 can be most accurately predicted when 3 nearest neighbors are used to classify the new cases. During the training and testing of KNN algorithm, available data is randomly split into two parts, with 80% training data and 20% testing data. Four separate KNN models are built to predict the four flags in sequential order. Table 3.4 presents the accuracy of flag generation. After the prediction of 4 flags, 22 logging data and 5 flags are used together to predict the NMR T_2 distribution. Out of 416 discrete depths, 354 randomly selected depths are used for training and 62 remaining depths are used for testing.

Table 3.4 Accuracy of flag generation using KNN models.

Flag	**2**	**3**	**4**	**5**
Right	**366**	**358**	**354**	**366**
Wrong	**50**	**58**	**62**	**50**

3.1.4 Fit T2 distribution with Gaussian distributions

Genty et al. (2007) found that NMR T_2 distribution can be fitted using three Gaussian distributions expressed as

$$f(T_2') = A\sum_{i=1}^{3}(\alpha_i)\, g_i(\mu_i, \sigma_i, T_{2i}') \text{...(3.2)}$$

where i is an index that identifies the Gaussian distribution, $T_2' = log(T_2)$, g_i is the probability distribution function of a Gaussian distribution with mean μ_i and standard deviation σ_i, α_i represents the proportion of pore volumes representing the constituent Gaussian distribution with respect to total pore volume such that $\alpha_1 + \alpha_2 + \alpha_3 = 1$, and

A is the amplitude parameter. In BPS, NMR T_2 distributions have either one or two peaks. I fit the T_2 distributions using the modified function expressed as

$$f(T_2') = \sum_{i=1}^{2} (\alpha_i)\, g_i(\mu_i, \sigma_i, T_2') \text{..(3.3)}$$

Compared to Equation 3.2, Equation 3.3 does not implement the amplitude parameter A and $\alpha_1 + \alpha_2 \neq 1$. Six parameters are required to fit the T_2 distribution response at each depth.

Furthermore, the reliability of the fitting is expressed in terms of correlation coefficient (R^2) formulated as

$$R^2 = 1 - {}^{RSS}\!/_{TSS} \text{...(3.4)}$$

$$RSS = \sum_{i=1}^{n} \left[f_{i,fit}(T_2') - f_i(T_2') \right]^2 \text{...(3.5)}$$

$$TSS = \sum_{i=1}^{n} \left[f_i(T_2') - \overline{f(T_2')} \right]^2 \text{...(3.6)}$$

where n = 64 is the number of discretized samples of the original T_2 distribution response of each depth, $f_{i,fit}(T_2')$ is the i-th discretized sample of the T_2 distribution response computed using Equation 3.3 and the six corresponding fitting parameters, $f_i(T_2')$ is the i-th discretized sample of the original T_2 distribution response, and $\overline{f(T_2')}$ is the mean of the original T_2 distribution. RSS is the sum of squares of the residuals and TSS is the total sum of squares proportional to the variance of the data. T_2 distribution responses acquired at 416 depth points in BPS were fitted with Equation 3.3 to estimate the characteristic six fitting parameters for each depth point. In doing so, six logs are generated that can be used for training and testing phases of the second predictive model. $\alpha_2 = \mu_2 = \sigma_2 = 0$ when there is only one peak in the T_2 distribution. Figure 3.6 shows the results of fitting for randomly sampled depth points. T_2 distributions were fitted at median R^2 of 0.983 (Figure 3.7) Only 12% of the depths were fitted with R^2 lower than 0.95.

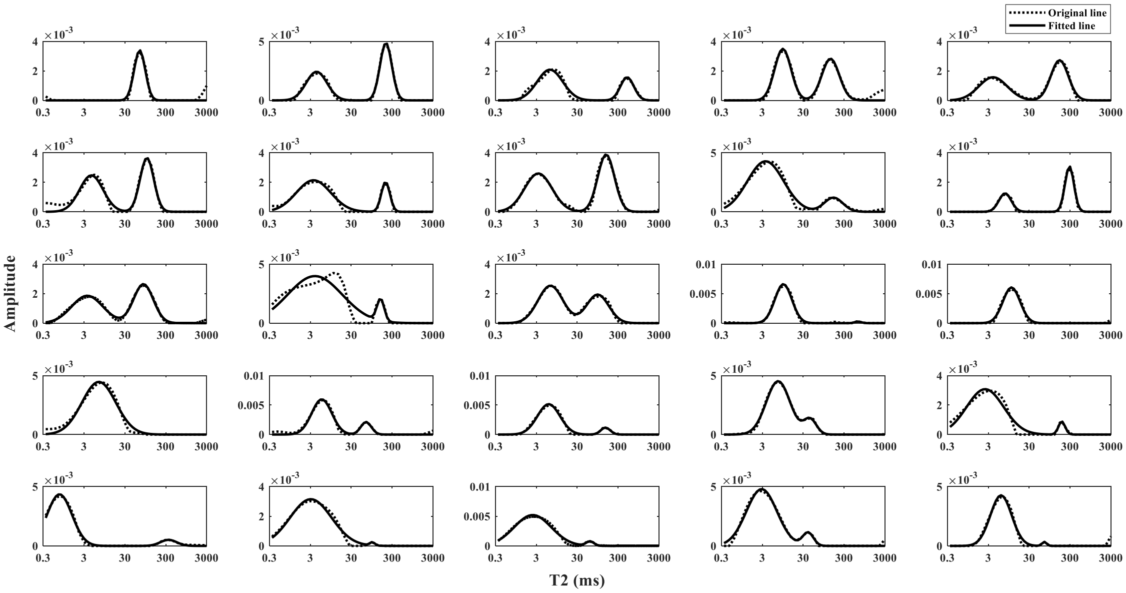


Figure 3.6 Dashed curves are original T2 distributions and solid curves are the best-fitting T2 distributions obtained using Equation 3.3, such that dashed curve identifies original and solid curve identifies fitted data.

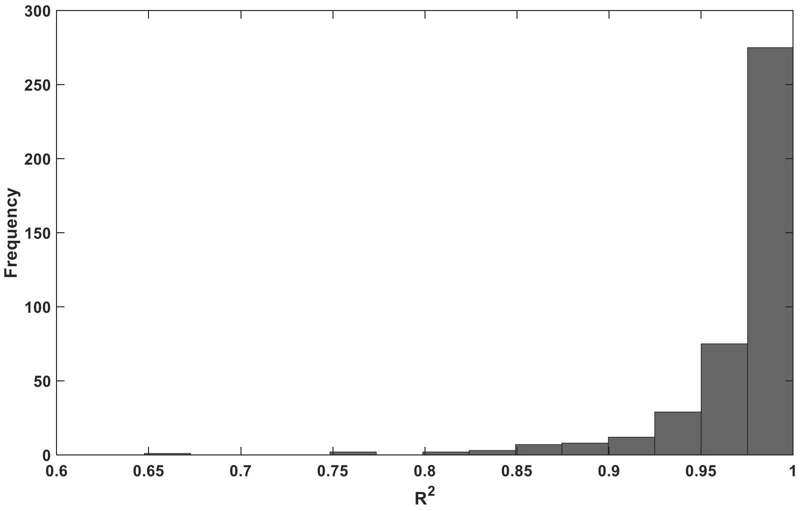


Figure 3.7 R^2 distribution of NMR T_2 distribution fitting results.

In addition, Normalized Root Mean Square Error (NRMSE) is used together with R^2 to assess the accuracy of NMR T_2 distribution prediction by ANN models. The equation of NRMSE is shown below

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \text{……………………………………………………..……........(3.7)}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \text{………………………………………………………….........(3.8)}$$

$$NRMSE = \frac{RMSE}{y_{max} - y_{min}} \text{…………………………………………………………...........(3.9)}$$

from the function, it is obvious that the range of NRMSE is from 0 to 1 and I will obtain precise prediction performance of my model when it is close to 0. NRMSE is opposite to R^2 when evaluating the accuracy since larger R^2 means higher accuracy.

3.1.5 Data Preprocessing

Before building an ANN model, data preprocessing is necessary to make inputs and outputs more suitable for the prediction of models. Few outliers are removed. Outliers are abnormally large or small compared to other data points. Their existence will change weights and bias of ANNs so that they have a negative effect on the accuracy of prediction. For example, at some depths, Gamma Ray can be larger than 1000 API unit, or DTS can be larger than 800 $\mu s/ft$, which is unrealistic. Furthermore, inputs and outputs are normalized so that the minimum values of each input and output are -1 and the maximum values are 1. Normalization forces inputs and outputs to lie in the same range, which guarantees stable convergence of weights and biases (Genty, 2006). Normalization was performed using the following equation

$$y = 2\frac{x - x_{min}}{x_{max} - x_{min}} - 1 \qquad (3.10)$$

where x is the original value of input or output and y is the normalized value. In addition, dataset is split into two parts: training data and testing data. Usually, 80% of data are selected as training data and the remaining 20% form the testing data. In my models, due to the limited nature of the available data, 85% of the dataset are randomly selected to be the part of training data and the remaining 15% form the testing data.

3.1.6 Build ANN models

Two ANN models are built following the data preprocessing step. Each ANN model is built with two hidden layers, which is enough for most function approximation problems. There are no specific equations to calculate the number of neurons in each hidden layer when the number of inputs and outputs are known. Different combinations of neurons in each layer were tried for the NMR T_2 prediction. An arithmetic sequence of the number of neurons in each hidden layer generates high prediction accuracy (Demuth at al. 2014). Consequently, for the first predictive model that takes 27 inputs and generates 64 outputs, 39 and 51 neurons were set in the first and second hidden layers of the ANN model, respectively, such that 27, 39, 51 and 64 is close to an arithmetic sequence. This architecture requires 6460 weights and biases to be computed during each training step. Following the same logic, for the second predictive model that has 27 inputs and 6 outputs, requires 20 and 13 neurons in the first and second hidden layers of the ANN model, respectively, such that the sequence 6, 13, 20 and 27 is nearly an arithmetic sequence. This architecture requires 917 weights and biases to be computed during each training step.

Several algorithms can be applied as training functions, which are models to adjust weights and biases in order to converge target functions of ANNs. Target functions are models to describe errors and it is the goal of ANNs to minimize them using training functions. Levenberg-Marquardt (LM) backpropagation (Chamjangali et al., 2007) and Conjugate Gradient (CG) backpropagation (Cheng at al., 2005) are two most widely used algorithms in ANN models for training function approximation problems, which build a relationship between inputs and outputs with weights and biases of every neuron in the ANN model. LM backpropagation is suitable for small number of weights and biases, whereas CG backpropagation can be applied for large neural networks implementing large number of weights and biases. When the number of weights and biases increases, the iteration speed of LM algorithm will decrease more than that of CG algorithm. CG backpropagation is applied as the training function of both the ANN models. Training time of the models with LM backpropagation was 10 times more than that with CG backpropagation in the first model. To be specific, scaled conjugate gradient algorithm (Cheng at al., 2005) is selected and tested as the best training function of the ANN models specific to the NMR T_2 prediction work.

In the target function of ANN models, the purpose is to adjust weights and biases of all neurons to minimize it so that the error of ANN models is the smallest to ensure best prediction performance. Overfitting is the main problem when minimizing the value of target function during iterations (Kuhn and Johnson, 2013). ANN models cannot recognize the appropriate relationship between inputs and outputs because of overfitting. One basic target function is the Sum of Squared Errors (SSE) function expressed as

$$\mathrm{SSE} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{P} \sigma_j^2 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(3.11)$$

where n is the number of samples (416 in this study) and P is the number of outputs (64 for the first model and 6 for the second model), λ is penalty parameter, y_i is original outputs, $\hat{y}_i$ is estimated outputs, and σ_j^2 are variances of outputs. Regularization model is utilized to introduce a penalty parameter in the target function to avoid overfitting. By sacrificing some bias, variance can be reduced to minimize SSE function (Kuhn and Johnson, 2013). Penalty parameter is set at 0.7 in the first model and 0.5 in the second model after trial and error. In other models, people also avoid overfitting by dividing some samples into validation data. There is no need to divide data as validation data in my models because of regularization. In consequence, more data can be used for training models to obtain more robust models.

3.2 Case Study

22 conventional logging data and 5 flags are used as inputs to predict 64 discretized amplitudes in the first model and 6 fitting parameters in the second model that characterize the T_2 distributions in BPS. Data from 416 different depths are randomly split into 354 (85%) depths to be used as the training data and 62 (15%) depths as the testing data. R^2 and NRMSE are used to evaluate the accuracy of T_2 predictions by ANN models. All 64 bin amplitudes predicted in the first model can be compared with the original discretized amplitudes to evaluate the prediction accuracy. For the second model, the predictions of 6 fitting parameters is used to reconstruct the T_2 prediction as a sum of two Gaussian distributions for each depth, and then compared with the original T_2 distribution to evaluate the prediction accuracy.

3.2.1 Training the first ANN Model

Figure 3.8 presents prediction performance for the training phase of the first ANN model using 25 randomly selected depths. Few peaks are not matched properly during the training phase that is indicative of noise in the input logs and the T_2 distributions used for training. Low prediction accuracy in Training phase will result in poor performance during testing.

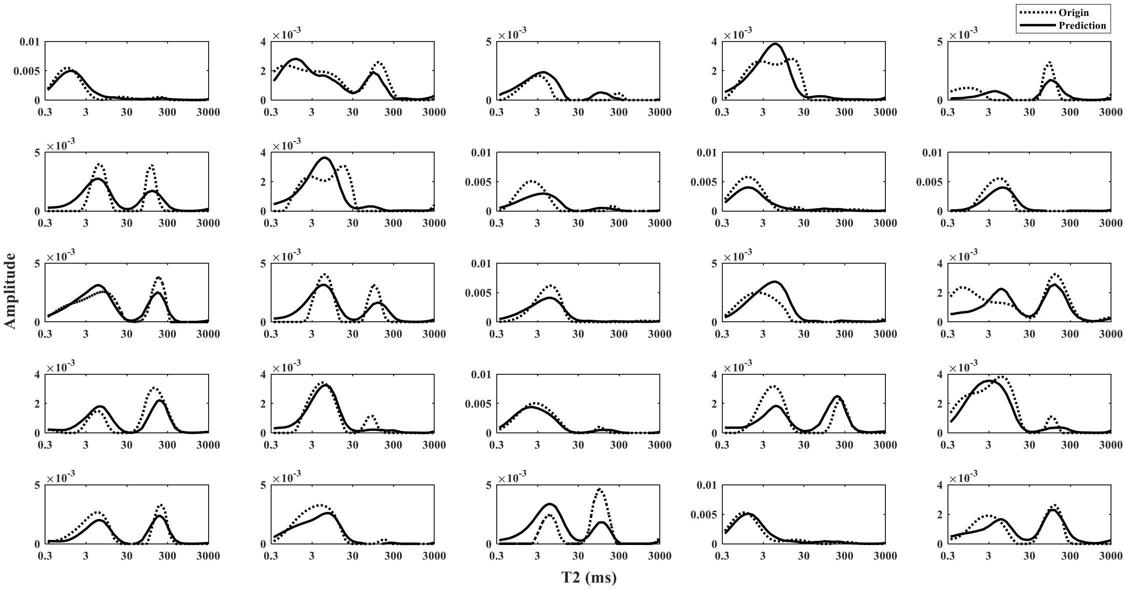


Figure 3.8 Comparison of original T2 distributions with those predicted using the first ANN model during the training phase, such that dashed curve identifies original and solid curve identifies predicted data.

The median R^2 of predictions during the training phase is 0.8574 and the median NRMSE is 0.1201, which indicate a good prediction performance. R^2 and NRMSE distributions for 354 depths are plotted together in Figure 3.9. R^2 can be negative according to Equation 3.4 if $RSS > TSS$, which indicates that the predicted trend does not follow the measured data. When R^2 is equal to 0 or negative, it means that the prediction performance of these predicted curves is worse than a mean-value fit (Mendenhall and Sincich, 2016). For the R^2 distributions presented in Figure 3.9, I include negative R^2 to the histogram at 0 so that all R^2 lie in [0, 1]. In contrast, all NRMSE lie in [0, 1]. During the training phase, most R^2 are larger than 0.7 and most NRMSE are smaller than 0.2 implying a good prediction performance of the first ANN model.

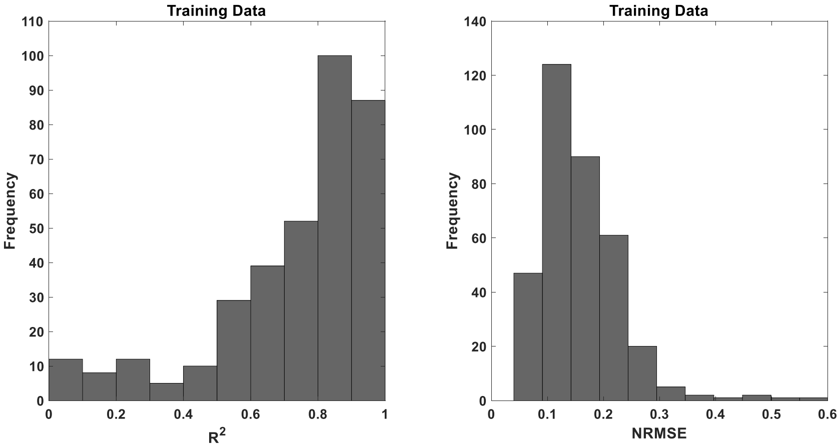


Figure 3.9 R^2 and NRMSE distributions of prediction performance using the first ANN model during the training phase.

Four examples of prediction performance for different values of R^2 are illustrated in Figure 3.10 to aid qualitative understanding of the training results. The first subplot at the top left with $R^2 = 0.99$ is the case with best performance. During the training, all depths with single-peak T_2 distribution are trained at a high prediction performance compared to those with two peaks. The subplot on top right with $R^2 = 0.85$ is the median performance case, and more than half of all depths will perform better than the one shown for $R^2 = 0.85$. $R^2 = 0.50$ is a bad prediction performance, and about 7% of all depths are trained at lower prediction performance than this one. The bottom right subplot with $R^2 = 0$ is a poor performance example. Less than 3% depths will perform as bad as the one shown for $R^2 = 0$.

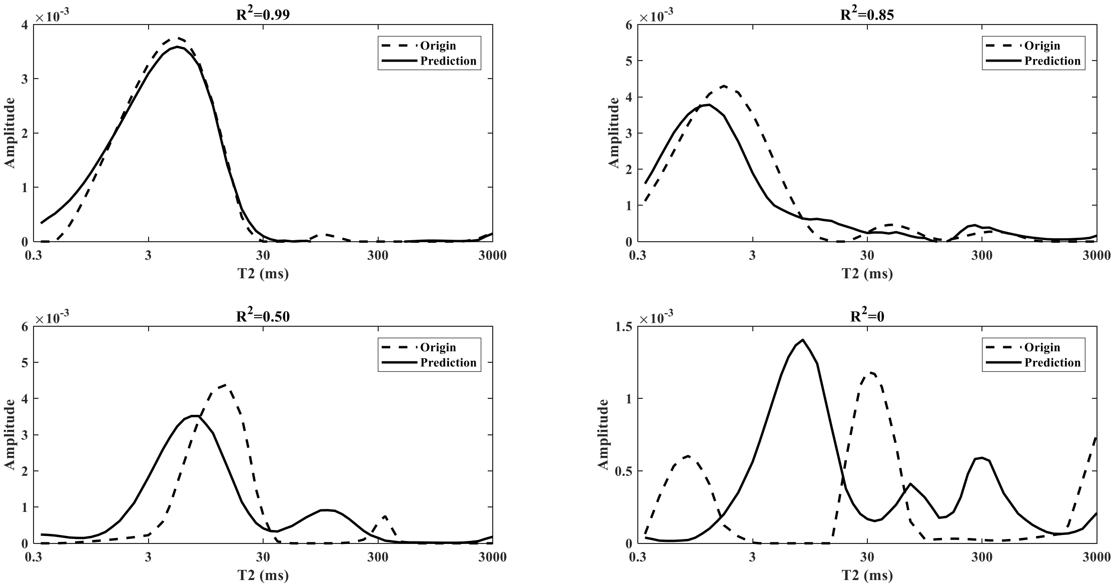


Figure 3.10 Prediction performance of the first ANN model for various R^2.

3.2.2 Testing the first ANN Model

There are only 62 testing depths, which are comparatively fewer than 354 depths for training. The prediction accuracy during the testing phase is similar to that attained during the training phase. The median R^2 of testing is 0.8549 and the median NRMSE is 0.1218 obtained after using noisy and limited inputs and outputs to train and test the ANN model. Figure 3.12 presents prediction performance for the testing phase of the first ANN model using 25 randomly selected depths from the testing depths. all testing depths have prediction performance higher than $R^2 = 0.6$ except for one case and about 90% of testing depths have prediction performance lower than NRMSE=0.2 (Figure 3.13).

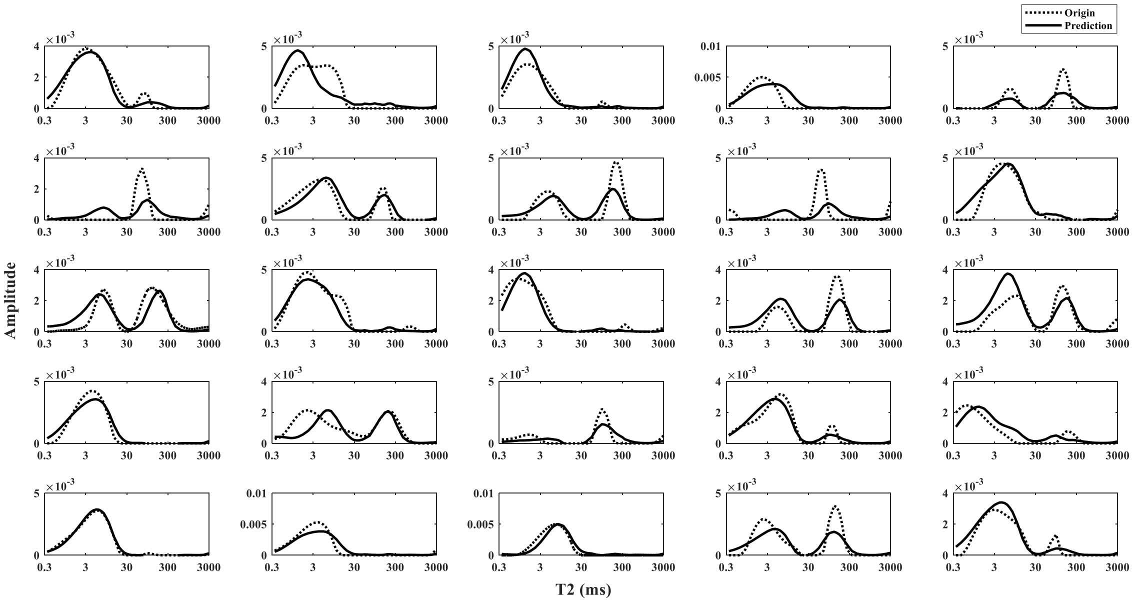


Figure 3.11 Comparison of original T2 distributions with those predicted using the first ANN model during the testing phase, such that dashed curve identifies original and solid curve identifies predicted data.

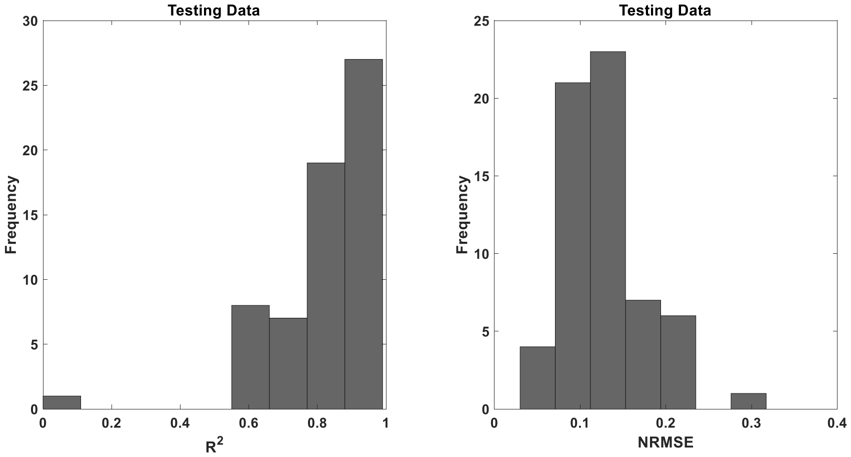


Figure 3.12 R^2 and NRMSE distributions of prediction performance using the first ANN model during the testing phase.

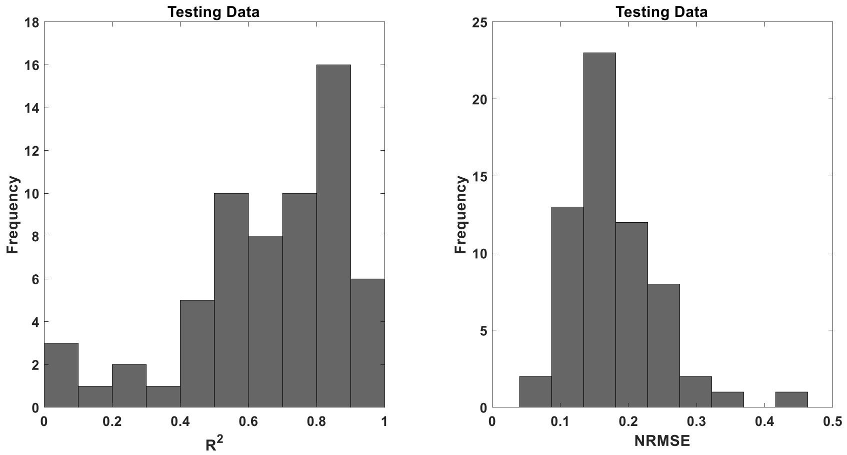


Figure 3.13 R^2 and NRMSE distributions of prediction performance using the first ANN model during the testing phase (without flags).

3.2.3 Training the second ANN Model

Prediction performance in second model is affected by the fact that prior to T_2 prediction six parameters need to be computed that characterize the T_2 distribution as sum of two Gaussian distributions. This introduces error prior to prediction.

The median R^2 of predictions during the training phase is 0.7634 and the median NRMSE is 0.1571 for the second ANN model, compared to 0.8574 and 0.1201, respectively, for the first ANN model. It is of high accuracy although only limited and noisy data are used in the model. Consequently, the prediction performance of the first ANN model is superior to that of the second model, but the computational time of the first ANN model is 1.3 times as compared to that of the second model. R^2 and NRMSE distributions for 354 depths are plotted together in Figure 3.14. During the training phase, most R^2 are larger than 0.7 and most NRMSE are smaller than 0.25 implying an acceptable prediction performance of the second ANN model. $R^2 = 0.5$ is a bad prediction performance, and about 22% of all depths are trained at lower prediction performance than this one, which is higher than that for the first model by 5%. Less than

5% depths will perform as bad as the one shown for $R^2 = 0$, which is also higher than that for the first model by 2%.

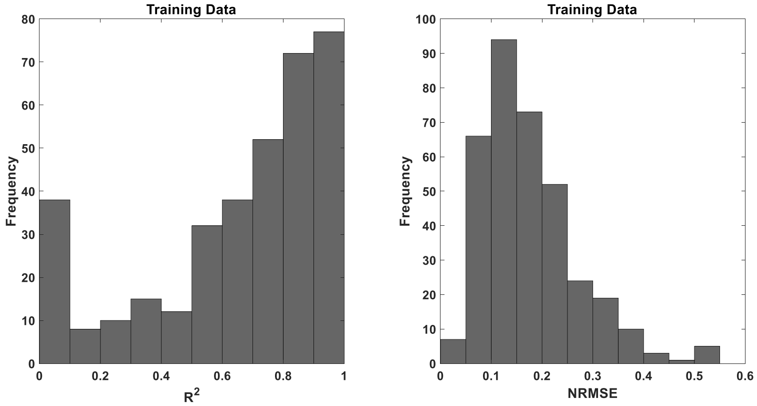


Figure 3.14 R^2 and NRMSE distributions of prediction performance using the second ANN model during the training phase.

3.2.4 Testing the second ANN Model

The prediction accuracy during the testing phase is similar to that attained during the training phase. The median R^2 of testing is 0.7584 and the median NRMSE is 0.1609, which is high accuracy with noisy and limited data used in the ANN model. 29% of testing depths have prediction performance lower than $R^2 = 0.5$ and 37% of testing depths have prediction performance higher than $R^2 = 0.8$ (Figure 3.15).

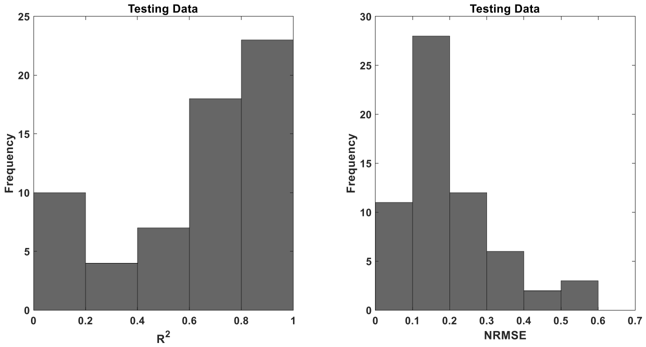


Figure 3.15 R^2 and NRMSE distributions of prediction performance using the second ANN model during the testing phase.

3.2.5 Validation of the first Model

NMR T_2 distributions are generally used to estimate the formation porosity and permeability, which are the two most important hydrocarbon-reservoir parameters. In this section, I derive few reservoir properties from the predicted and original NMR T2 distribution to test the robustness of the predicted NMR T2. The first model is validated by comparing ϕ_N and $T_{2,gm}$ derived from the original NMR T_2 distribution with those derived from the synthetically generated T2 distribution. ϕ_N is the sum of all amplitudes of the 64 bins of a T2 distribution at a single depth, $T_{2,gm}$ is the 64-th root of the product of the 64 discretized T_2 amplitudes at a single depth. Schlumberger-Doll-Research (SDR) model is a popular model for the estimation of permeability based on ϕ_N and $T_{2,gm}$, which is expressed as

$$k_{SDR} = C * T_{2,gm}^2 * \phi_N^4 = C * \left(T_{2,gm}\phi_N^2\right)^2 \text{..}(3.11)$$

where k_{SDR} is the permeability computed using the SDR model. I derived the SDR-model term, $T_{2,gm}\phi_N^2$, in Equation 3.11 from the original and predicted NMR T2 distribution and then compare them to test the prediction performance of the permeability estimation using the SDR model on the generated NMR T2. Comparison results are presented in Figure 3.16. Table 3.5 indicates that the NMR T2 generated using the first predictive method can be used to compute the three reservoir parameters of interest with good accuracy.

Table 3.5 Accuracy of ϕ_N, $T_{2,gm}$ and $T_{2,gm}\ \phi_N^2$ derived from the synthetically generated NMR T_2 distribution.

	$\boldsymbol{\phi_N}$	$\boldsymbol{T_{2,gm}}$	$\boldsymbol{T_{2,gm}\phi_N^2}$
$\boldsymbol{R^2}$	**0.7685**	**0.8664**	**0.7587**
NRMSE	**0.0909**	**0.0840**	**0.0854**

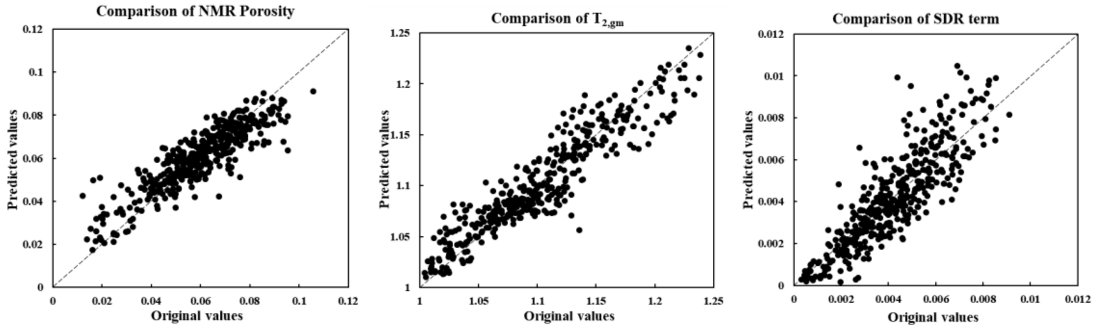


Figure 3.16 Comparison of ϕ_N, $T_{2,gm}$ and $T_{2,gm}\ \phi_N^2$ computed from the original NMR T_2 distributions against those computed from the synthetically generated NMR T_2 distributions.

3.2.6 Importance Ranking of Inputs

The importance of every input is also ranked by replacing them one by one with 0 and train the model again to evaluate the accuracy drop. Large accuracy drop without the information of one input means that this input is of much significance to the model. In Figure 3.17, flags 1, 2 and 3 are ranked at the top and flags 4 and 5 are ranked in the middle level, which proves the necessity of designing these flags. These flags are all very important for NMR T_2 distribution because they are designed according to pore size distributions. For example, if only the lithology flag is removed, there will be an accuracy drop of about 3.5%.

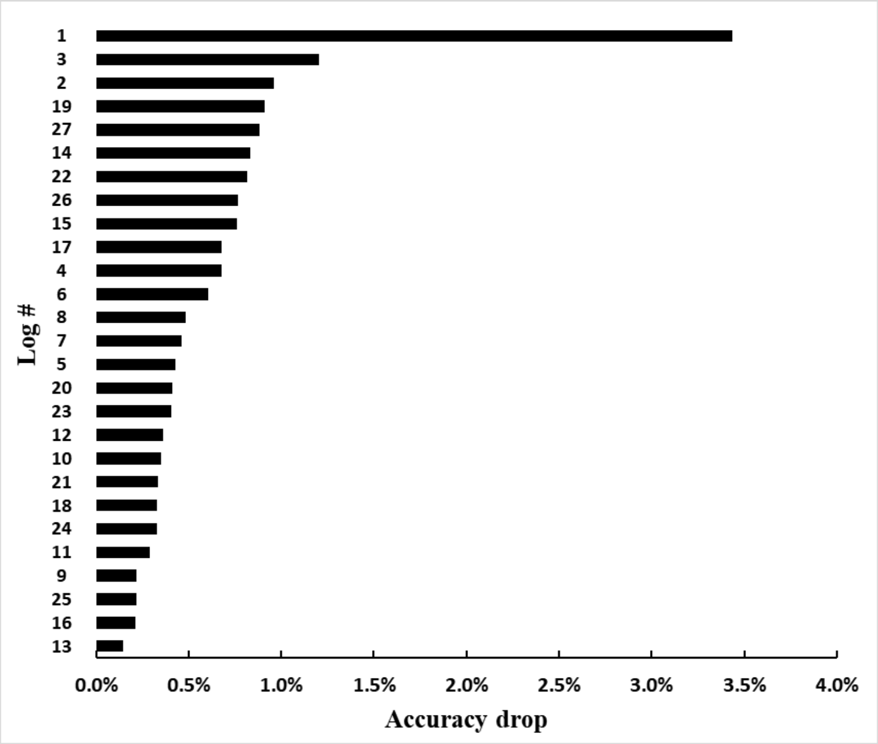


Figure 3.17 Ranking importance of inputs in the first ANN model (1-5: flags1-5, 6: GR, 7: DPHZ, 8: NOPR, 9: PEFZ, 10: RHOZ, 11: VCL, 12: AT10, 13: AT90, 14: DTC, 15: DTS, 16: VPVS, 17: Caliper, 18: Illite, 19: Chlorite, 20: Bound water, 21: Quartz, 22: K-Feldspar, 23: Clacite, 24: Dolomite, 25: Anhydrite, 26:Unflushed water, 27: Unflushed oil).

3.3 Conclusions

22 logs and 5 qualitative flags were processed by two distinct neural network models to generate the NMR T_2 distribution responses, which approximate the in-situ fluid-filled pore size distribution in BPS. The first predictive model generates T_2 distribution discretized into 64 bin amplitudes, whereas the second predictive model generates the 6 fitting parameters that characterize the T_2 distribution as a sum of two Gaussian distributions. The first predictive model performs more accurately, exhibiting median R^2 of 0.8549 during testing, compared to the second one, exhibiting median R^2 of 0.7584. However, the second model has lower computational cost compared to the first model. Input data used in my predictive models were limited in quantity and prone to

noise and uncertainty due to the subsurface borehole conditions in which they were acquired. Nonetheless, the two predictive models exhibit good prediction performance. A few reservoir properties, ϕ_N, $T_{2,gm}$ and $T_{2,gm}\phi_N^2$, were derived from the synthetic T_2 distribution at reasonable accuracies.

Chapter 4: Generation of Synthetic Dielectric Dispersion Logs in Organic-Rich Shale Formations Using Neural-Network Models

4.1 Theory and Methodology

In this thesis, conventional triple combo logs, sonic logs, and DD logs are acquired in three wells. Wells 1 and 2 are in PB formation in Two Georges field, Texas and Well 3 is in BS formation in Hawkeye field, North Dakota. Data preparation and data preprocessing are performed prior to the application of the three predictive methods illustrated in Figures 4.1, 4.2 and 4.3, respectively. Four discrete frequencies at which DD logs were acquired in PB formation were 20 MHz, 100 MHz, 300 MHz and 1 GHz and those in BS formation were 22 MHz, 100 MHz, 250 MHz, and 0.96 GHz. The first predictive method implements an ANN model that processes conventional log data to simultaneously predict the 8 DD logs, namely four conductivity and four relative permittivity logs at four discrete frequencies in the range of 10 MHz to 1 GHz (Figure 4.1). The second predictive method involves a two-step process (Figure 4.2) in which the four conductivity dispersion logs are simultaneously predicted using one ANN model. Subsequently, a second ANN model processes the four previously predicted conductivity logs with the conventional logging data to simultaneously predict the four permittivity dispersion logs. The third predictive method based on proprietary technique performs the best both in predicting conductivity and permittivity dispersion logs. It cannot be disclosed in the thesis due to patentable materials. We will demonstrate the improvements in the prediction performance when using the third method.

I developed and compared the three methods to demonstrate the effects of architecture of neural networks on the generative capabilities. The three predictive methods use different arrangements of neural networks, inputs, outputs, and number of neurons in hidden layers. General practitioner uses neural network architecture similar to that implemented in the first predictive method involving simultaneous generation. - Multivariate Linear Regression (MLR) model, which is the classical regression model, can also be used for such generative work. In Well 1, the prediction performance of the MLR model is 9.6% worse for conductivity dispersion and 11.2% worse for permittivity dispersion than that of the first predictive method, which is the worst performing method out of the three methods studied in this thesis. The first method performs simultaneous generation using a single ANN and serves as the baseline for the prediction performance.

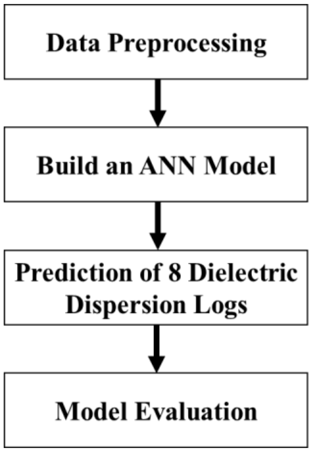


Figure 4.1 Flowchart for the first predictive method.

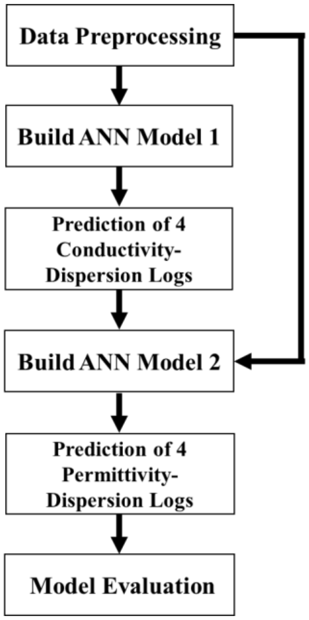

Data Preprocessing
Build ANN Model 1
Prediction of 4 Conductivity-Dispersion Logs
Build ANN Model 2
Prediction of 4 Permittivity-Dispersion Logs
Model Evaluation

Figure 4.2 Flowchart for the second predictive method.

4.1.1 Data Preparation

The predictive methods are applied to following depth intervals: (a) 2200 ft in Well 1 intersecting 6 different intervals in PB formation and (b) 1300 ft in Well 2 intersecting 4 different intervals in PB formation, containing shale, siltstone and sandstone; and (c) 500 ft intersecting 7 different intervals in BS formation containing shale, siltstone, sandstone and dolostone. Gamma ray log (GR), density porosity log (DPHZ), neutron porosity log (NPOR), photoelectric factor log (PEFZ), bulk density log (RHOZ), volume of clay log (VCL), delta-T compressional sonic log (DTC), delta-T shear sonic log (DTS) and laterolog resistivity logs at 6 depths of investigation (RLA0, RLA1, RLA2, RLA3, RLA4, RLA5) are selected as the easy-to-acquire conventional logs that are fed into the three predictive models. These conventional logs (Tracks 2-5, Figure 4.3) and the DD logs (Tracks 7 & 8, Figure 4.3) are used to train and test the predictive methods. DD logs comprise 4 conductivity dispersion logs (Track 7) and 4 permittivity dispersion logs (Track 8) acquired at four discrete frequencies in the frequency range of 10 MHz to 1 GHz.

The main criteria for selecting the easy-to-acquire logs chosen for the proposed prediction are as follows: (a) the easy-to-acquire logs and the 8 DD logs should be influenced by similar petrophysical properties, and (b) the easy-to-acquire logs should be acquired in most of the wells as a part of the conventional logging plan. For example, I use resistivity logs as inputs because the properties influencing resistivity also affect permittivity and conductivity dispersions of a material, such that the DD logs and resistivity logs are influenced by porosity, saturation, volume of clay and pyrite, invasion, presence of interfacial polarization, texture of pore surfaces, tortuosity, and pore topology. There may be redundancies in the 15 “easy-to-acquire” conventional logs, for example, DPHZ and NPOR. However, I performed extensive sensitivity analysis and found that the absence of any one of the l5 logs lowers the prediction performance. Redundancies and dependencies among the input logs lowers the adverse effects of noise in the logs. I observe a 3% improvement in accuracy when using both the DPHZ and NPOR logs.

One synthetic discrete-valued log is computed in each well. This flag indicates lithology/formation-type of an interval. Flag improves the prediction accuracy. Flag assumes an integer between [1, 6] in Well 1, [1, 4] in Well 2, and [1, 7] in Well 3 depending on the numbers of distinct lithology or formation type intersected by the well. Flag log can be developed based on cores, offset logs, logs, seismic boundaries, clustering algorithms, or knowledge about the reservoir. For the first predictive method, the prediction performances without the flag in Well 1 is 8.0% worse in conductivity dispersion and 9.3% worse in permittivity dispersion than those with the flag. In summary, there are 15 conventional log inputs, including one Flag log, and 8 output logs. Depth corrections were performed on all the logs prior to the processing.

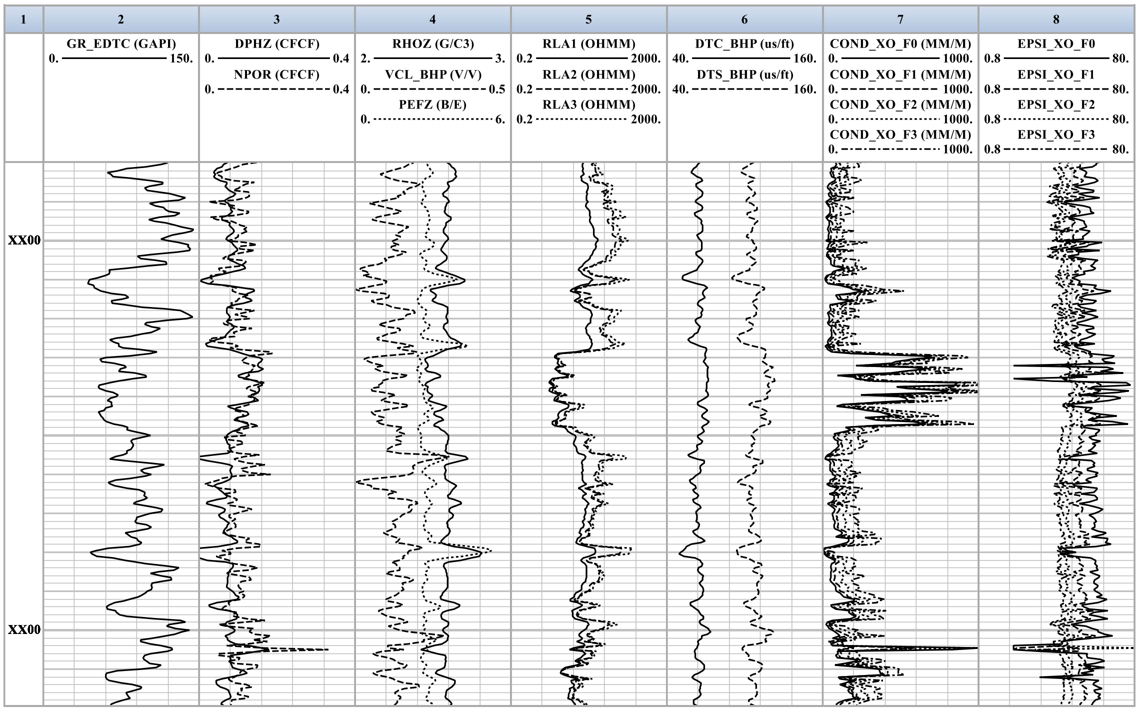


Figure 4.3 Track 1 is depth, Track 2 is gamma ray log, Track 3 contains density porosity and neutron porosity logs, Track 4 contains formation photoelectric factor, bulk density and volume of clay logs, Track 5 is laterolog resistivity logs at 3 depths of investigation (RLA1, RLA2, RLA3), Track 6 contains DTC and DTS logs, Track 7 is shallow conductivity dispersion logs, and Tracks 8 is shallow permittivity dispersion logs.

4.1.2 Measurement of Prediction Performance

Two parameters are used to compare the prediction performances of the predictive methods during the training and testing phases. The first parameter is correlation coefficient (R^2), which is formulated as

$$R_j{}^2 = 1 - RSS_j/TSS_j \quad \text{...(4.1)}$$

where

$$RSS_j = \sum_{i=1}^{n} \left(D_{pi,j} - D_{mi,j} \right)^2 \text{……………………………………………....................}(4.2)$$

and

$$TSS_j = \sum_{i=1}^{n} \left(D_{pi,j} - \overline{D}_j \right)^2 \text{…………………………………………………….............}(4.3)$$

such that n is the number of depths for which prediction needs to be performed, j = 1, 2, 3, and 4 indicates the four conductivity dispersion logs and j = 5, 6, 7, and 8 indicates the four permittivity dispersion logs, $D_{pi,j}$ is the ε_r or σ response predicted at depth i, $D_{mi,j}$ is the log j (ε_r or σ) response measured at depth i, and $\overline{D}_j$ is the mean of log j (ε_r or σ) responses measured at all depths for which training or testing is being performed. RSS_j is the sum of squares of the residuals and TSS_j is the total sum of squares proportional to the variance of the corresponding log j responses. As a result, R^2 will be affected by the variance of data. If TSS_j is small, ${R_j}^2$ is more likely to be small, which represents a bad prediction performance and it may not be true. When most data are close to the mean value, which represents low variance of data, even good prediction performance can predict a low ${R_j}^2$ (Vapnik, 2013). In conclusion, ${R_j}^2$ is not very suitable for estimating accuracy when measured logs have low variance.

The second parameter is Normalized Root Mean Square Error (NRMSE) formulated as

$$RMSE_j = \sqrt{\frac{\sum_{i=1}^{n}(D_{pi,j} - D_{mi,j})^2}{n}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4.4)$$

$$NRMSE_j = \frac{RMSE_j}{D_{mi,j,max} - D_{mi,j,min}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4.5)$$

High prediction accuracy is indicated by NRMSE less than 0.1. NRMSE measures the error with respect to the range of data. Using NRMSE, the percentage of error to the range of data can be known, which facilitates better assessment of accuracy in comparison to RMSE. If there are some data much larger or smaller than the rest, they need to be deleted as outliers in the data preprocessing step. In doing so, $D_{mi,j,max} - D_{mi,j,min}$ cannot attain large inconsistent values. When outliers are not removed, NRMSE will disguise as a low prediction performance despite large RMSE. In comparison between R^2 and NRMSE, R^2 will not be trustworthy when measuring data of low variance, whereas NRMSE is trustworthy for all cases once outliers are removed.

4.1.3 Data Preprocessing

Before building an ANN model, data preprocessing is necessary to make inputs and outputs more suitable for purposes of prediction. On an average, 2% of depth points used for training and testing exhibited outlier responses and were removed during data preprocessing. For example, at few depths, Gamma Ray responses were close to 1000 API units and for few DTS responses were higher than 800 $\mu s/ft$. Data from such depths will change weights and bias of ANN models; thereby having an adverse effect on the prediction performance. Following that, input and output logs were normalized so that all data are within [-1, 1].

80% of data are selected as training data randomly and the remaining 20% form the testing data in three wells. During supervised learning, dataset can be divided into three parts for training, validating and testing phases. In this thesis, I divide the data into training and testing dataset. Validation data reduces overfitting. Instead of using a validation dataset, I use a penalty parameter to prevent overfitting. Validation data is not required when using the penalty parameter, as a result more data is available for the training and testing phases.

4.1.4 Neural Network Architectures

For the first method, only one ANN model is built with 15 inputs and 8 output DD log predictions. For the second method, the first ANN model contains 15 inputs and 4 output conductivity dispersion log predictions and the second ANN model contains 19 inputs, including the 4 predicted conductivity dispersion logs, and 4 output permittivity dispersion log predictions. An arithmetic sequence of the number of neurons in each hidden layer typically predicts a high prediction accuracy (Demuth et al., 2014). For example, 12 and 10 neurons are used in the first and second hidden layers, respectively, for the ANN model implemented in the first method. In the second predictive method, 11 and 7 neurons are used in the first and second hidden layers, respectively, of the first ANN model, whereas 14 and 9 neurons are used in the first and second hidden layers, respectively, of the second ANN model.

CG backpropagation is applied as the training function after comparison of training time with Levenberg-Marquardt (LM) backpropagation algorithm. Training time of the ANN models with LM backpropagation was 2-5 times more than that with CG backpropagation. To be specific, scaled CG backpropagation (Cheng et al., 2005) is selected and tested as the best training function for the ANN models implemented in this thesis.

Overfitting is the main problem when minimizing the value of target function during iterations. ANN model cannot recognize the appropriate relationship between inputs and outputs because of overfitting. One basic target function is Sum of Squared Errors (SSE) function, which is expressed as

$$\mathrm{SSE} = \sum_{i=1}^{n}(D_{mi,j} - D_{pi,j})^2 + \lambda \sum_{j=1}^{P} \sigma_j^2 \quad \text{...(4.6)}$$

where n is the number of depths in a well, j = 1, 2, …, 8 represents the 8 output DD logs, P=8 is the total number of output logs, λ is the penalty parameter, $D_{mi,j}$ is the value of measured DD log j at depth i, $D_{pi,j}$ is the value of predicted output log j at depth i, and σ_j^2 is the variance of the predicted output log j, such that

$$\sigma_j^2 = \frac{1}{n}\sum_{i=1}^{n}(D_{pi,j} - \mu_j)^2 \quad \text{...(4.7)}$$

where μ_j is the mean of the predicted output log j. Regularization method is utilized to introduce a penalty parameter in the target function to avoid overfitting and ensure a balanced bias-variance tradeoff. λ ranges from 0 to 1, which is set based on extensive numerical experiments. For example, it is 0.1 for the ANN in the first method, 0.05 for the two ANNs in the second method, and 0.2, 0.3, 0.2, 0.15, 0.3, 0.2, 0.2, and 0.2 for the 8 ANNs in the third method. For the training in a single well, the first method requires on an average 5 seconds. Computational cost of the second method is twice that of the first method, whereas the computational cost of the third methodis 7-8 times that of the first method.

4.2 Case Study

I compare prediction performances of the three predictive methods by applying them to three wells, with Wells 1 and 2 in PB formation and Well 3 in BS formation. There are several petrophysical differences between the two formations. Water salinity is about 360,000 ppm in BS (Simpson, 2015) and less than 50,000 ppm in PB (McNeal, 1965). As a result, water salinity from Well 3 in BS is much higher than that from Wells 1 and 2 in PB. There are 4293, 2456, and 920 depth points selected in Well 1, Well 2, and Well 3, respectively, for purposes of training and testing the neural network models.

4.2.1 Prediction Performance of the first Predictive Method

In Well 1, I test the performances of ANN models with and without NMR logs as inputs. 859 depth points in Well 1 were used for testing the trained ANN model of the first method in Well 1. When using NMR logs as additional inputs, I include 8 NMR-bin porosity logs and NMR T_2-logarithmic mean log. With NMR logs as inputs, 24 inputs and 8 outputs are required to train and test the ANN model in the first predictive method. Following the principle of arithmetic sequence in number of neurons in an ANN model, 18 and 13 neurons are used in the first and second hidden layers, respectively. As shown in Table 4.1, there is no improvement in the prediction accuracy when using NMR logs as additional conventional log inputs. This is counterintuitive because both NMR and DD logs are sensitive to porosity and pore-filling fluids. However, DD logs sense storage and conductance of electric charges in pore-filling fluid, whereas NMR logs sense mobility and relaxation of H1 nuclei. Most likely, this difference between two logs impedes the improvement of prediction performance. Consequently, all predictions from here on do not use NMR logs as inputs. Another noticeable feature is prediction performances of conductivity dispersion logs are better than permittivity dispersion logs, which may be a consequence of larger noise in permittivity dispersion logs due to tool physics and inversion methodologies and greater sensitivity of the permittivity dispersion logs to interfacial polarization mechanisms. Well 2 intersects a geological formation that is very similar to the one intersected by Well 1. However, in comparison to Well 1, prediction accuracy of first method is lower in Well 2 (Table 4.1) due to fewer samples available for the training in Well 2, which was a 43% reduction in number of samples. Interestingly for both the wells, the prediction performances for conductivity dispersions decrease with increase in frequency.

The first predictive method fails in DD log prediction in Well 3 intersecting the BS formation (Table 4.1). In Well 3, the prediction performance of the first method is 90.6% worse in conductivity dispersion and 65.6% worse in permittivity dispersion in terms of NRMSE, compared with that in Well 1. One reason for bad prediction performance of the first method in Well 3 is the number of samples. There are 4293, 2456, and 920 depth points selected in Well 1, Well 2, and Well 3, respectively. For Well 2, there is only about 57% of data compared to Well 1 that led to the decrease of prediction accuracy (Table 4.1). Depth points in Well 3 are even fewer than those in Well 2, which leads to low prediction performance in Well 3. Another possible reason for the poor prediction performance in Well 3 is the high water salinity in BS formation. From electromagnetic measurement standpoint, increase in salinity predominantly increases transport of electrical energy and also increases the storage of electromagnetic energy due to interfacial polarization phenomena. Among the conventional logs, resistivity logs are highly sensitive to water salinity due to the enhancement of charge transport. However, the increase in salinity masks the physical relationship between ε_r and water content and enhances the effects of interfacial polarization mechanisms on permittivity dispersions, which do not have as high a physical correlation with other conventional logs. As a result, high water salinity masks the physical relationships between DD logs and the conventional logs resulting in the poor prediction performance in Well 3. Interestingly, the prediction performances of conductivity dispersion are always better than those of permittivity dispersion because conductivity dispersion is tied to the charge transport resulting in stronger physical relationships with conventional logs, especially resistivity logs, than permittivity dispersion. In addition, for most inputs and outputs in Well 3, S_d/μ and skewness are higher than those in Wells 1 and 2 (Table 4.2). S_d/μ, referred as the coefficient of variation, is the standard deviation over mean value of data that represents the relative spread and skewness is a measure of the asymmetry of data. High values of both indicate that the logs in Well 3 are either noisier or the formations exhibit more geological/petrophysical variability compared to those in Wells 1 and 2. This will also lead to poor prediction performance in Well 3. Notably, the prediction performances of permittivity dispersion logs increase with increase in frequency, unlike Wells 1 and 2.

Table 4.1 Comparison of prediction performances of the first predictive method in Wells 1, 2 & 3 with and without NMR logs.

		NRMSE				R^2			
		f0	**f1**	**f2**	**f3**	**f0**	**f1**	**f2**	**f3**
Well 1 without NMR	**Conductivity**	**0.056**	**0.063**	**0.074**	**0.087**	**0.93**	**0.92**	**0.90**	**0.86**
	Permittivity	**0.096**	**0.112**	**0.091**	**0.097**	**0.70**	**0.71**	**0.65**	**0.62**

Well 1 with NMR	**Conductivity**	**0.069**	**0.077**	**0.091**	**0.098**	**0.90**	**0.89**	**0.86**	**0.82**
	Permittivity	**0.104**	**0.097**	**0.088**	**0.086**	**0.62**	**0.66**	**0.60**	**0.55**
Well 2 without NMR	**Conductivity**	**0.075**	**0.081**	**0.098**	**0.104**	**0.91**	**0.90**	**0.85**	**0.81**
	Permittivity	**0.136**	**0.150**	**0.140**	**0.141**	**0.67**	**0.67**	**0.63**	**0.52**
Well 3 without NMR	**Conductivity**	**0.150**	**0.128**	**0.129**	**0.102**	**0.71**	**0.76**	**0.72**	**0.76**
	Permittivity	**0.192**	**0.175**	**0.154**	**0.133**	**0.04**	**0.18**	**0.33**	**0.54**

Table 4.2 Variances and skewnesses of input and output logs, Sd/μ is the standard deviation over mean value of data and skewness is a measure of the asymmetry of data.

	S_d/μ			Skewness		
	Well 1	**Well 2**	**Well 3**	**Well 1**	**Well 2**	**Well 3**
Interval	**0.477**	**0.466**	**0.286**	**-0.050**	**-0.143**	**-1.192**
GR	**0.348**	**0.281**	**1.466**	**0.473**	**-0.121**	**3.745**
DPHZ	**0.597**	**0.578**	**3.050**	**-0.413**	**-0.388**	**0.834**
NPOR	**0.558**	**0.416**	**0.852**	**0.069**	**-0.144**	**0.393**
PEFZ	**0.249**	**0.168**	**0.170**	**1.368**	**0.602**	**0.235**
RHOZ	**0.030**	**0.015**	**0.051**	**0.413**	**0.327**	**-0.833**
VCL	**0.536**	**0.405**	**0.699**	**0.095**	**-0.043**	**-0.005**
RLA0	**0.063**	**0.044**	**0.205**	**0.555**	**-0.726**	**3.235**
RLA1	**0.594**	**0.497**	**0.735**	**-0.010**	**-0.449**	**4.433**
RLA2	**1.080**	**0.989**	**1.023**	**1.967**	**1.418**	**1.573**
RLA3	**1.303**	**1.314**	**1.586**	**2.079**	**2.593**	**1.915**
RLA4	**1.969**	**1.468**	**2.104**	**4.969**	**3.553**	**0.157**
RLA5	**2.159**	**1.496**	**2.073**	**6.495**	**3.805**	**0.797**
DTC	**0.134**	**0.078**	**0.210**	**-0.290**	**-0.436**	**1.994**
DTS	**0.101**	**0.713**	**0.162**	**0.110**	**-12.604**	**1.302**
σ_{f0}	**1.207**	**0.837**	**1.277**	**1.307**	**1.264**	**1.073**
σ_{f1}	**1.037**	**0.703**	**1.361**	**1.131**	**1.068**	**0.980**
σ_{f2}	**0.863**	**0.616**	**1.112**	**0.879**	**0.775**	**1.056**

$\boldsymbol{\sigma_{f3}}$	**0.747**	**0.577**	**0.968**	**0.677**	**0.561**	**0.717**
$\boldsymbol{\epsilon_{r,f0}}$	**0.393**	**0.452**	**1.066**	**0.346**	**0.299**	**2.497**
$\boldsymbol{\epsilon_{r,f1}}$	**0.271**	**0.253**	**0.607**	**0.392**	**0.448**	**2.366**
$\boldsymbol{\epsilon_{r,f2}}$	**0.208**	**0.165**	**0.496**	**0.464**	**0.668**	**2.057**
$\boldsymbol{\epsilon_{r,f3}}$	**0.151**	**0.096**	**0.372**	**0.175**	**0.482**	**1.023**

4.2.2 Prediction Performance of the second Predictive Method

The second method is applied to Wells 1 and 2. Based on the NRMSE, the second method performs better in predicting permittivity dispersion logs compared to the first method. In second method, the first ANN model is first trained to predict the conductivity dispersion without the constraint of matching the permittivity dispersions; consequently, there is some sacrifice in the prediction accuracy for conductivity dispersion, which are 7.6%, 10.4%, 9.3% and 6.8% higher for the frequencies f0, f1, f2, and f3, respectively, in terms of NRMSE when compared to the first method (Table 4.3). Higher NRMSE represents lower accuracy. However, the second ANN model is trained to predict permittivity dispersion using the measured conductivity dispersion. In doing so, there is an improvement in the accuracy of predicting permittivity dispersion logs, which are 7.8%, 10.7%, 3.1% and 3.4% lower for f0, f1, f2, and f3, respectively, in terms of NRMSE when compared to the first method. In comparison to permittivity dispersions, conductivity dispersions have stronger physical relationships with conventional logs, especially with resistivity logs. Therefore, I first predict conductivity dispersion (Table 4.3), following that the predicted conductivity dispersion is used to predict the permittivity dispersion. Owing to the low variance of $\epsilon_{r,f3}$ log, R^2 is not a good indicator for evaluating prediction accuracy of permittivity dispersion.

Table 4.3 Comparison of prediction performances of the second predictive method in Wells 1 & 2.

		NRMSE				R^2			
		f0	f1	f2	f3	f0	f1	f2	f3
Well 1 Second method	Conductivity	0.063	0.075	0.086	0.093	0.94	0.92	0.89	0.85
	Permittivity	0.095	0.090	0.088	0.090	0.65	0.69	0.59	0.49
Well 2 Second method	Conductivity	0.078	0.084	0.102	0.111	0.90	0.89	0.83	0.78
	Permittivity	0.119	0.144	0.136	0.140	0.59	0.56	0.49	0.33

4.2.3 Prediction Performance of the third Predictive Method

In comparison to first method, the prediction performance of the third method for permittivity dispersion decreases by 9.1%, 13.3%, 5.6%, and 5.9% for f0, f1, f2, and f3, respectively, in terms of NRMSE (Table 4.4). Prediction accuracy for σ_{f0} drops and the rest increases in the third method compared to the first and second methods. Further, in comparison to first method, the prediction performance of the third method for conductivity dispersion changes by 6.1%, -0.7%, -4.1%, and -4.8% for f0, f1, f2, and f3, respectively, in terms of NRMSE (Table 4.4). All three predictive methods have decent accuracy although the dataset is prone to noise and is limited in terms of sample size.

Table 4.4 Comparison of prediction performances of the third predictive method in Wells 1 & 2.

		NRMSE				R^2			
		f0	f1	f2	f3	f0	f1	f2	f3
Well 1 Third method	Conductivity	0.067	0.066	0.071	0.077	0.92	0.92	0.90	0.87
	Permittivity	0.093	0.088	0.089	0.086	0.69	0.72	0.66	0.62
Well 2 Third method	Conductivity	0.072	0.077	0.094	0.105	0.92	0.91	0.87	0.83
	Permittivity	0.118	0.139	0.129	0.138	0.71	0.70	0.65	0.52

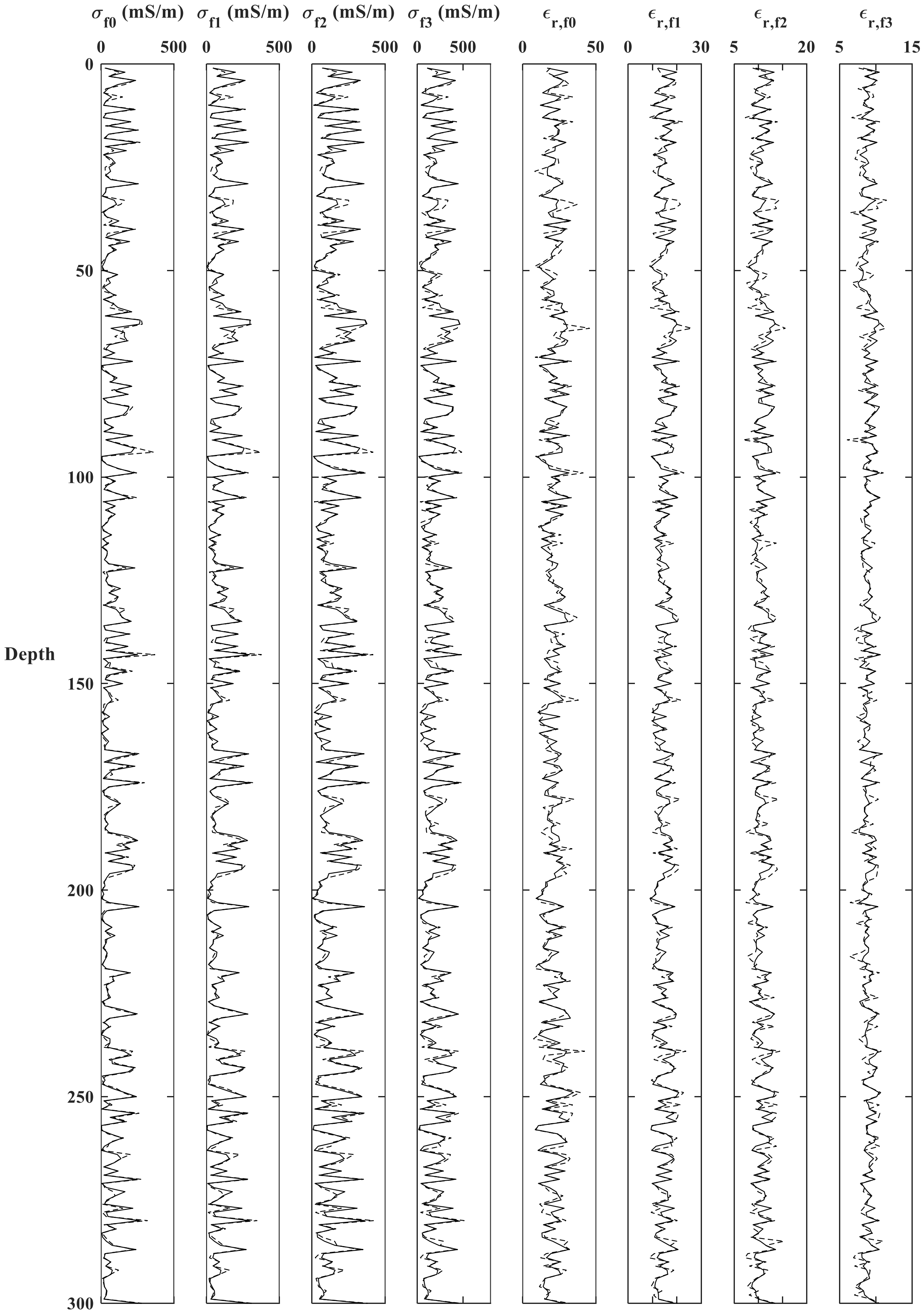


Figure 4.4 Comparison of the 8 original (dashed) dielectric dispersion logs with those predicted (solid) using the third prediction method in Well 1 (1-300 testing data).

4.2.4 Application of the third Predictive Method Trained and Tested in Well 1 to Well 2 and that in Well 2 to Well 1

In the first case, 8 ANN models implemented in the third method are trained and tested in Well 1. Following that the third predictive method is deployed in Well 2. In the second case, the training and testing is performed in Well 2 and then deployed in Well 1. For both the cases, prediction performances of the third method when trained/tested and deployed in separate wells (Table 4.5) are lower than those obtained when trained and deployed in the same well, as described in Section 3.3. In comparison to the third method applied in Well 2, the prediction performances of the third method trained in Well 1 and deployed in Well 2 for conductivity dispersion logs change by 8.3%, 19.5%, 0% and -3.8% and those for permittivity dispersion logs change by 5.1%, -5.0%, 15.5%, 2.2% for f0, f1, f2, and f3, respectively, in terms of NRMSE. In Wells 1 & 2, the formations have similar sequence of intervals but the thicknesses of the intervals vary between the two wells. On an average, prediction performance for conductivity dispersion logs drops by 6% and permittivity dispersion logs drops by 4.5% compared to those obtained when the predictive method is trained and deployed in the same well. However, the results presented in this section show the feasibility of training the third predictive method in one well and then deploying it in another well for predicting DD logs.

Table 4.5 Comparison of prediction performances of the third predictive method when trained/tested and deployed in separate wells.

		NRMSE				R^2			
		f0	**f1**	**f2**	**f3**	**f0**	**f1**	**f2**	**f3**
Method trained and tested in	**Conductivity**	**0.078**	**0.092**	**0.094**	**0.101**	**0.89**	**0.86**	**0.85**	**0.80**

Well 1 and deployed in Well 2	Permittivity	0.124	0.132	0.131	0.141	0.67	0.67	0.63	0.48
Method trained and tested in Well 2 and deployed in Well 1	Conductivity	0.103	0.092	0.105	0.116	0.85	0.84	0.82	0.79
	Permittivity	0.124	0.131	0.141	0.138	0.66	0.65	0.60	0.55

4.2.5 Sensitivity Analysis

For purposes of the sensitivity analysis, first I study the importance of the 15 conventional logs, including the one synthetic discrete-valued log denoting the lithology type. NRMSEs for the generation of each conductivity and permittivity dispersion logs are summed together to compute the accuracy drop. Sensitivity of the ANN model for simultaneous generation to the conventional input logs are reported in Figure 4.5. ANN model is the most sensitive to the removal of sonic (DTC & DTS), neutron porosity (NPOR), discrete-valued lithology flag, medium resistivity (RLA3), and clay volume (VCL) log. ANN model is the least sensitive to the removal of shallow resistivity (RLA1 and RLA2), rock density logs (RHOZ), and photoelectric factor (PEFZ). There are 6 resistivity logs that sense different reservoir volumes, such that RLA0 is shallowest and RLA5 is deepest sensing resistivity log. Figure 4.5 indicates that the DD log generation is the most affected when RLA3 is removed and the least affected with RLA2 is removed. Resistivity logs should be the most important when predicting DD logs because of physical relations between resistivity, conductivity and permittivity. However, none of the resistivity logs with different depth of investigations were ranked high (Figure 4.5) because of the high correlations between the various resistivity logs. In other words, no matter which resistivity log is removed to test its importance, the other five will provide sufficient information preventing the accuracy drop. The accuracy drop of removing all resistivity logs together is about 10.2% indicating highest dependence of dielectric dispersion logs on resistivity logs.

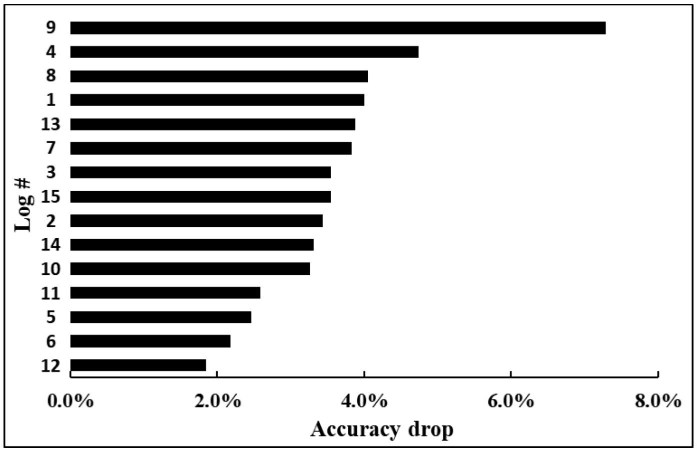


Figure 4.5 Comparison of reduction in prediction performance when one of the 15 conventional logs is removed one at a time. Each integer on y-axis identifies the removal of a specific log, such that 1: Lithology Flag; 2: GR; 3: DPHZ; 4: NPOR; 5: PEFZ; 6: RHOZ; 7: VCL; 8: DTC; 9: DTS; 10-15: Resistivity at different depths of investigation (10: RLA0; 11: RLA1; 12: RLA2; 13: RLA3; 14: RLA4; 15: RLA5).

After identifying the importance of each input log for the proposed log generation, I determine the smallest set of inputs required for maintaining the desirable prediction accuracy. This is determined by deleting inputs one by one starting with the least important input (Figure 4.6). I found that at least 11 most important inputs should be retained to maintain an accuracy drop less than 10%. This smallest set of 11 log inputs is obtained by removing RLA2, RHOZ, PEFZ and RLA1 (Figure 4.7). As shown in Figure 4.7 (bottom-most bar), a set of inputs containing the 6 most important inputs results in 17% drop in prediction accuracy. Notably, when the 6 resistivity logs of different depths of investigation are retained and other logs are deleted one by one based on their importance, I observe accuracy drop less than 10% upon the deletion of the 6 least important logs, namely RHOZ, PEFZ, GR, DPHZ, VCL and Lithology. I conclude that the correlations of inputs with outputs and the correlations among the inputs control the prediction performance.

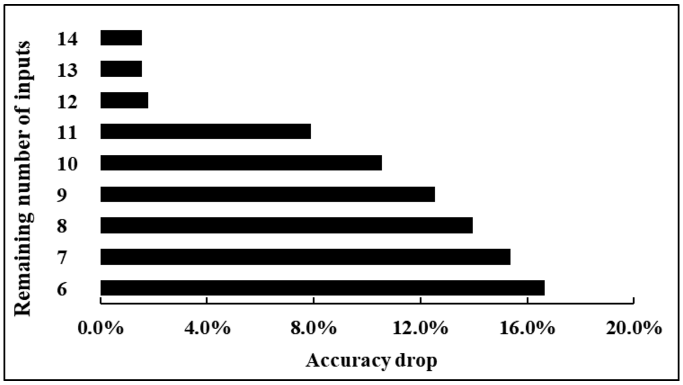


Figure 4.6 Comparison of reduction in prediction performance by deleting inputs one by one based on the importance of an input for the proposed log generation.

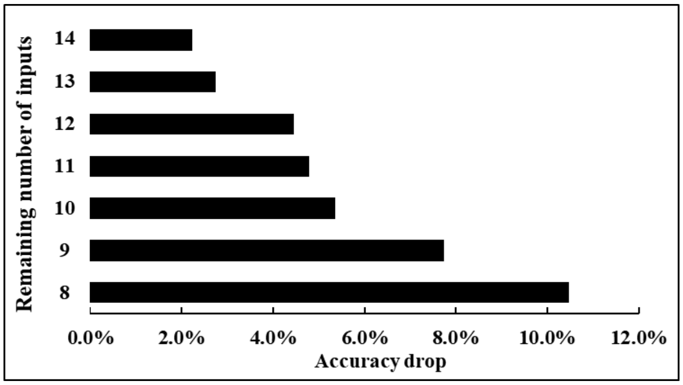


Figure 4.7 Comparison of reduction in prediction performance by deleting inputs (other than the 6 resistivity logs) one by one based on the importance of an input for the proposed log generation.

To study the sensitivity of the prediction performance to noise in training/testing data, 20% Gaussian noise is added one a time to each conventional log input, to the 6 resistivity log inputs together (i.e. 3.33% noise is added to each resistivity log), and to 8 DD log outputs together (i.e. 2.5% noise is added to each DD log). The sensitivity to noise in input log is ranked from top to bottom: Resistivity, DTS, GR, RHOZ, NPOR, VCL, DPHZ, DTC, and PEFZ (Figure 4.8). 20% of overall noise in the six resistivity logs results in maximum reduction in the prediction performance by 4.5%. Prediction performance is also highly sensitive to noise in GR, DTS, and DD logs. Sensitivity to noise in data and that to input itself follows similar trends for few logs, e.g. PEFZ, which is least important for the log generation.

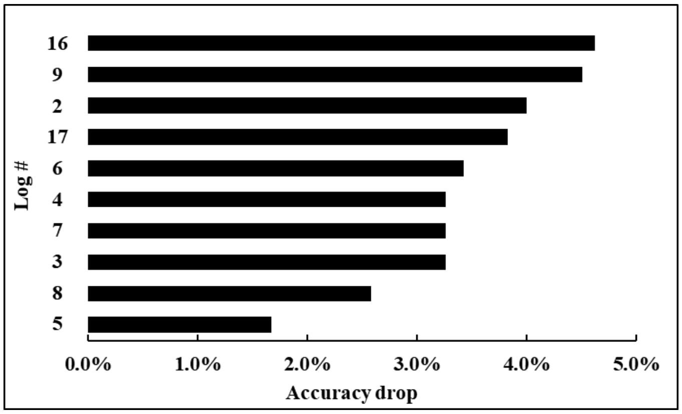


Figure 4.8 Comparison of reduction in prediction performance when 20% Gaussian noise is added to inputs and outputs one at a time. Integer-valued log indices are similar to those listed in Figure 4.7. Log #16 represents 6 resistivity logs together and Log #17 represents the 8 dielectric dispersion logs together.

4.3 Conclusions

Three predictive methods are developed to process 15 conventional logs and generate 8 synthetic DD logs, comprising 4 conductive dispersion and 4 permittivity dispersion logs. The first method simultaneously predicts the 8 DD logs. The second method first processes the 15 conventional logs to predict the 4 conductivity dispersion logs, which are then processed along with the 15 conventional logs to predict the 4 permittivity dispersion logs. In comparison to the first method, prediction performance of the second method is 8.5% worse for conductivity dispersion and 6.2% better for permittivity dispersion. The third proprietary method exhibits the best predictive performance for the generation of the 8 synthetic DD logs. In comparison to the first method, prediction performance of the third method is 0.8% better for conductivity dispersion and 8.5% better for permittivity dispersion. Performances of these models are adversely affected by the noise in the logs and the limited amount of data available for the training purposes. Training the third predictive method in one well and then deploying it in another well for generating the 8 DD logs is feasible, such that the NRMSE of conductivity dispersion logs drops by 6% and that of permittivity dispersion logs drops by 4.5% compared to the baseline.

Chapter 5: Comparative Study of Shallow Learning Models for Sonic Log Prediction

5.1 Data Preparation and Processing

5.1.1 Data Preparation

In this thesis, six shallow learning models are applied to a shale reservoir in PB. Gamma ray log (GR), caliper log (DCAL), density porosity log (DPHZ), neutron porosity log (NPOR), photoelectric factor log (PEFZ), bulk density log (RHOZ), and laterolog resistivity logs at 6 depths of investigation (RLA0, RLA1, RLA2, RLA3, RLA4, RLA5) are selected as the easy-to-acquire conventional logs fed into the six models. Those input logs (Tracks 2-6, Figure 5.1) and DTC and DTS logs (Track 7, Figure 5.1) are used to train and test these models. One synthetic discrete-valued log is obtained from the lithology information of specific depth intervals. The formation is divided into several intervals according to lithology. The flag improves the prediction accuracy by setting an integer between [1, 13] along the well of 4240-feet. In summary, 13 inputs and 2 output logs were used for the training and testing purposes.

There are two main criteria for selecting the logs as inputs: (1) the input logs and sonic logs should be influenced by similar petrophysical properties, and (2) the easy-to-acquire input logs should be available in most of the wells as a part of the conventional logging plan. For example, DPHZ and NPOR are two logs measuring the porosity of the reservoir, which will affect DTC and DTS logs as well, and GR and DCAL are implemented in almost every well.

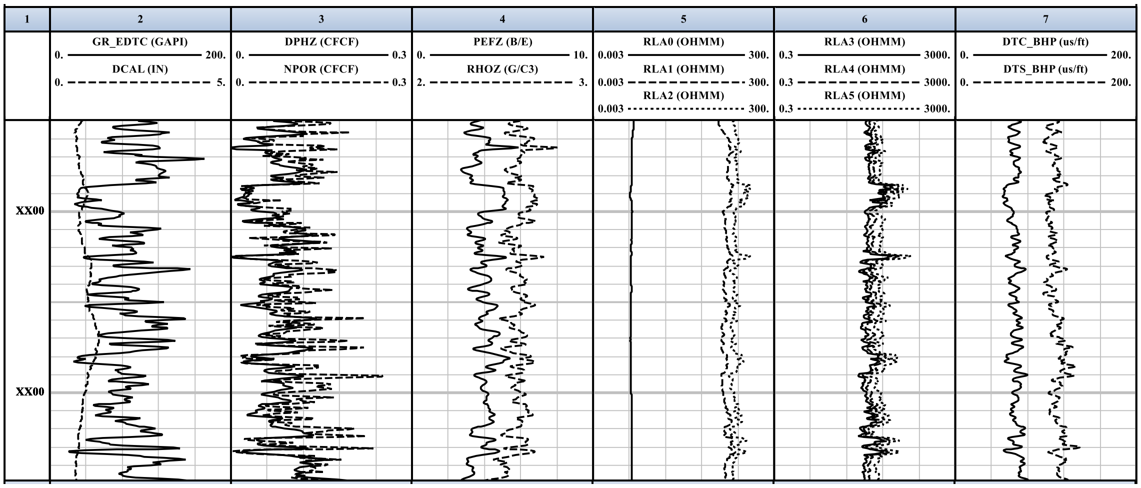


Figure 5.1 Track 1 is depth, Track 2 contains gamma ray and caliper logs, Track 3 contains density porosity and neutron porosity logs, Track 4 contains formation photoelectric factor and bulk density logs, Track 5 is laterolog resistivity logs at shallow depths of investigation (RLA0, RLA1, RLA2), Track 6 is laterolog resistivity logs at deep depths of investigation (RLA3, RLA4, RLA5), Track 7 contains DTC and DTS logs.

5.1.2 Data Preprocessing

Before using logging data to train models, data preprocessing is necessary to make input and output logs more suitable for purposes of prediction. No obvious outliers are detected in logging data. Normalization is necessary to transform all the data to the range of [-1, 1].

The dataset is split into two parts: training data and testing data. 80% of data are randomly selected as training data and the remaining 20% form the testing data for all models investigated in this thesis. The correlation coefficient (R^2) is used to compare the prediction performance of all models.

5.2 Methodology

Six shallow learning models are implemented for a comparative study of their prediction performances for generating the DTC and DTS logs by processing easy-to-acquire conventional logs. The first four models are linear regression models with different learning algorithms while the last two models are non-linear models with different structures.

5.2.1 Ordinary Least Squares (OLS) Model

OLS model is a linear regression model which optimizes the parameters of every input for the linear function (Draper and Smith, 2014). Parameters are calculated by minimizing the Sum of Squared Errors (SSE) between original outputs and predicted outputs in the model. The function of OLS model is formulated as

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (5.1)$$

where y_i are original outputs of the model, i is an integer ranging from 1 to n indicating distinct depth points, n is the number of observations (depth points along the length of the well, in this case), x_i are inputs of the model, p is the number of inputs (easy-to-accquire conventional logs, in this case), β are parameters of the OLS model, and ε is the error term. β and ε are determined by the OLS algorithm in the model. The loss function SSE is formulated as

$$\mathrm{SSE} = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 = \sum_{i=1}^{n} \varepsilon_i^2 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (5.2)$$

$$\hat{y}_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (5.3)$$

where $\hat{y}_i$ are predicted outputs of the model. Smaller SSE leads to better prediction performance. OLS model is such a simple model that people can use it for well log prediction easily and quickly. However, due to its simplicity, it also has some weaknesses. OLS model will be unduly influenced by outliers if they cannot be deleted before model construction. SSE is sensitive to outliers because all errors will be squared. Besides, it is impossible for the OLS model to detect high correlation between inputs and reduce dimensions before the model construction, which adversely affects the computational time.

5.2.2 Partial Least Squares (PLS) Model

PLS model is also a regression model, which constructs linear combinations of original inputs as new ones for prediction in order to reduce dimension (Draper and Smith, 2014). It is a supervised technique designed specifically for regression. New inputs combined with original ones are called components. The number of components m (no more than inputs) are chosen to maximally summarize the covariance with outputs. The determination of m build a simple PLS model with fewer inputs than OLS model when there are high correlations of inputs between each other. The flowchart of PLS algorithm is designed in Figure 5.2.

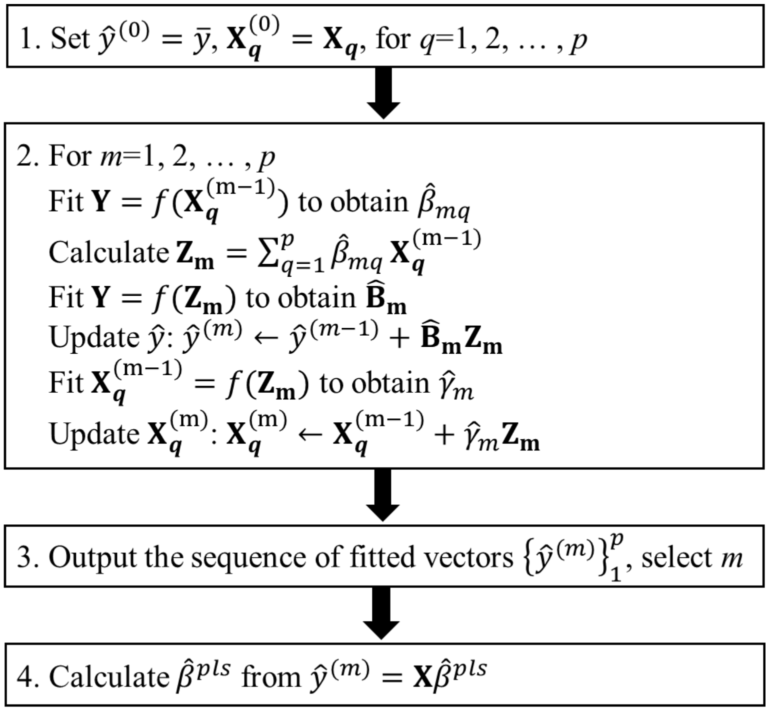

1. Set $\hat{y}^{(0)} = \bar{y}$, $\mathbf{X}_{\boldsymbol{q}}^{(0)} = \mathbf{X}_{\boldsymbol{q}}$, for q=1, 2, … , p
2. For m=1, 2, … , p
Fit $\mathbf{Y} = f(\mathbf{X}_{\boldsymbol{q}}^{(\mathrm{m}-1)})$ to obtain $\hat{\beta}_{mq}$
Calculate $\mathbf{Z}_{\mathbf{m}} = \sum_{q=1}^{p} \hat{\beta}_{mq} \mathbf{X}_{\boldsymbol{q}}^{(\mathrm{m}-1)}$
Fit $\mathbf{Y} = f(\mathbf{Z}_{\mathbf{m}})$ to obtain $\widehat{\mathbf{B}}_{\mathbf{m}}$
Update $\hat{y}$: $\hat{y}^{(m)} \leftarrow \hat{y}^{(m-1)} + \widehat{\mathbf{B}}_{\mathbf{m}}\mathbf{Z}_{\mathbf{m}}$
Fit $\mathbf{X}_{\boldsymbol{q}}^{(\mathrm{m}-1)} = f(\mathbf{Z}_{\mathbf{m}})$ to obtain $\hat{\gamma}_m$
Update $\mathbf{X}_{\boldsymbol{q}}^{(\mathrm{m})}$: $\mathbf{X}_{\boldsymbol{q}}^{(\mathrm{m})} \leftarrow \mathbf{X}_{\boldsymbol{q}}^{(\mathrm{m}-1)} + \hat{\gamma}_m \mathbf{Z}_{\mathbf{m}}$
3. Output the sequence of fitted vectors $\{\hat{y}^{(m)}\}_1^p$, select m
4. Calculate $\hat{\beta}^{pls}$ from $\hat{y}^{(m)} = \mathbf{X}\hat{\beta}^{pls}$

Figure 5.2 Flowchart of the PLS algorithm.

In Figure 5.2, X is the input matrix, Y is the original output matrix, m is the number of components, $\hat{\beta}$ is the parameter of the model, $\boldsymbol{Z_m}$ is the iteration term, $\hat{B}$ *and* $\hat{\gamma}$ are the iteration parameters. In the PLS algorithm, $\hat{\beta}$ and $\hat{y}$ are optimized for every m and the prediction performance will be measured and recorded for comparison. Smallest m with the best prediction performance is selected in the model. Sonic logs are predicted in the PLS model with constructed components, which are combinations of original inputs.

PLS model is good at dealing with inputs with high correlations because it can decrease the number of inputs after combination. It will also outperform other regression models in some cases where the number of inputs is larger than that of observations (Draper and Smith, 2014). However, it may lose useful information for prediction when constructing components from inputs, which leads to worse prediction performance than other models when there are no apparent redundant inputs.

In this case, Figure 5.3 shows the result of tuning the number of components. When the number of components is equal to 13, the best prediction performance is obtained. The results show that there are no obvious redundant inputs in the PLS model. Theoretically, m=13 and the prediction accuracy is the same as that of OLS model. In addition, when m=8, the prediction accuracy of PLS model will drop to 0.80 and 0.78 for DTC and DTS prediction respectively, in terms of R2.

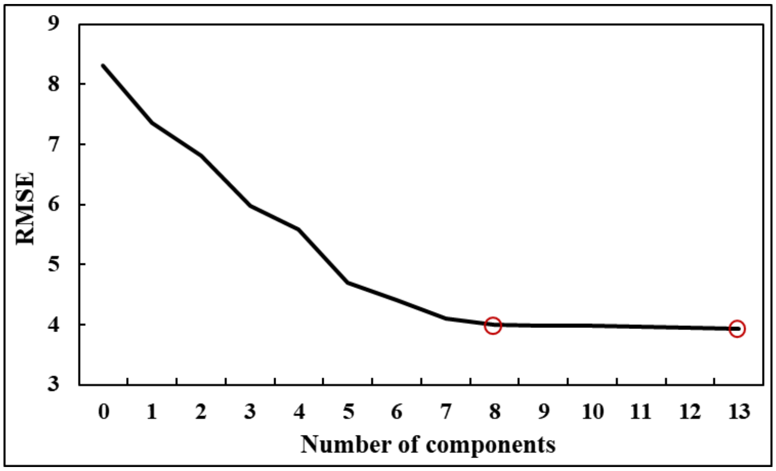


Figure 5.3 Tuning the parameter of number of components in the PLS model.

5.2.3 Least Absolute Shrinkage and Selection Operator (LASSO) Model

LASSO model is a regression model combined with a penalty term λ (Kuhn and Johnson, 2013), which is introduced to constrain the number of nonzero parameters. The loss function of LASSO model is similar to SSE, which is formulated as

$$L = \sum_{i=1}^{n}(y_i - x_i^T\beta)^2 + \lambda\sum_{q=1}^{p}\left|\beta_q\right| = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda\sum_{q=1}^{p}\left|\beta_q\right| \text{……..........(5.4)}$$

As λ increases, the number of nonzero parameters of β will decrease. LASSO model is designed for redundant inputs as well. larger λ in the loss function results in more zero parameters of β, which will eliminate the negative effect of redundant inputs when building the model. So LASSO model is similar to PLS model in dealing with large quantity of inputs. Unlike PLS model, which construct new components for better prediction performance, LASSO model just sets zero parameters to some inputs, which is easier to be understood and implemented. However, when there are no redundant inputs, LASSO model can fail to fully use all inputs for the best prediction performance unless λ is tuned to be 0.

λ is tuned from large to small numbers. Decreasing λ will increase $\sum_{q=1}^{p}\left|\beta_q\right|$ in the model (Figure 5.4). In this case, when λ is smaller than 4.83, $\sum_{q=1}^{p}\left|\beta_q\right|$ will increase significantly. So λ is optimized to be 4.83. For $\lambda = 4.83$, the prediction accuracy of LASSO model is 0.79 and 0.75 for DTC and DTS prediction respectively, in terms of R2.

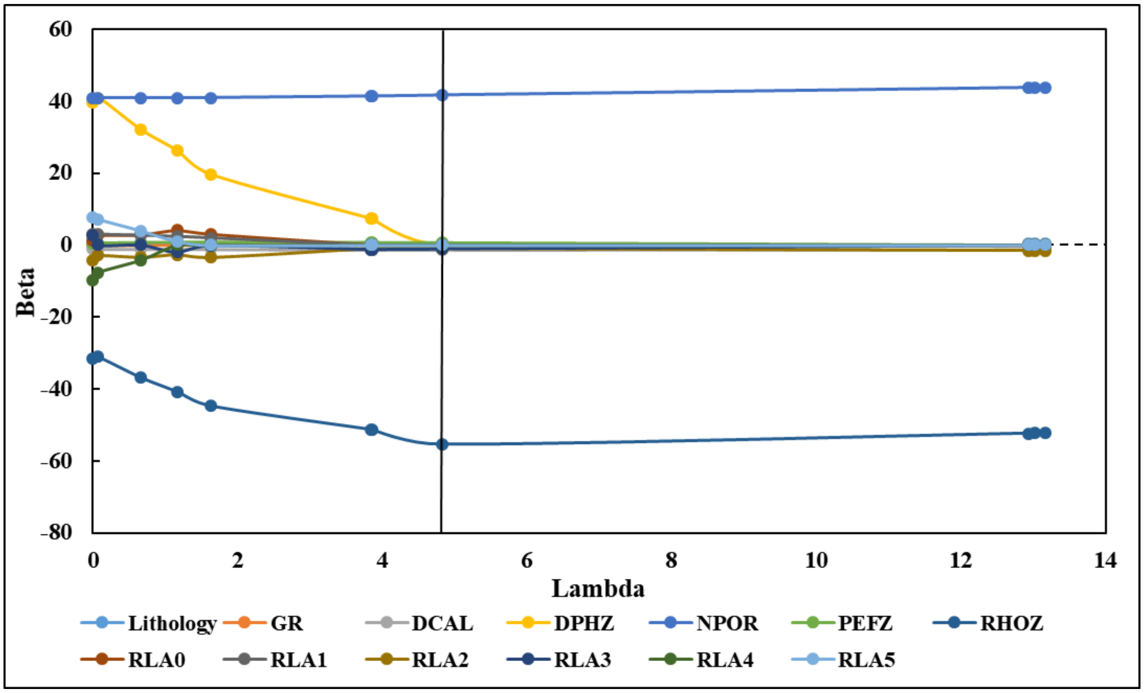


Figure 5.4 Tuning the parameter of λ in the LASSO model.

5.2.4 ElasticNet Model

ElasticNet model is a generalization of LASSO model (Kuhn and Johnson, 2013), which introduces two penalty terms (λ_1 and λ_2) in the loss function

$$L = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda_1 \sum_{q=1}^{p}\left|\beta_q\right| + \lambda_2 \sum_{q=1}^{p}\beta_q^2 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(5.5)$$

ElasticNet model is a stronger model than LASSO model for eliminating the effect of redundant inputs because there are two penalty terms in the loss function. Two parameters should be tuned when building the model. Similar to the LASSO model, it may not fully use all inputs for prediction when there are no redundant inputs in the case. It can also fail to make full use of all inputs for better prediction performance when there are no redundant inputs in the case.

In this case, λ_1 is tuned to be 4.8, which is close to λ in the LASSO model and λ_2 is tuned to be 0.1, which is very small. The prediction accuracy of ElasticNet model is 0.79 and 0.75 for DTC and DTS prediction respectively, in terms of R2. When there is no need to set stronger penalties in the loss function ($\lambda_2 \approx 0$), LASSO and ElasticNet will be close. When stronger penalties are needed in the loss function for a large number of redundant inputs. ElasticNet model will outperform LASSO model.

5.2.5 Multivariate Adaptive Regression Splines (MARS) Model

The first four models are all linear regression models. MARS model is a regression model but it is not based on simple linear combinations of inputs (Kuhn and Johnson, 2013). MARS model can fit separate linear regressions for different ranges with every input. The slopes as well as the number and range of the separate regions are estimated in the model. The combination of linear regressions for different ranges with all inputs will provide a curve for every predicted output, which takes non-linearities into consideration. The function of MARS model is formulated as

$$\hat{y} = \sum_{q=1}^{p} \alpha_q B_q(x_q) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(5.6)$$

where $B_q(x_q)$ are hinge functions and α_q are parameters of $B_q(x_q)$. A hinge function has the form of $\max(0, \mathrm{x} - \mathrm{C})$ and $\max(0, \mathrm{C} - \mathrm{x})$, where C is called knot of hinge functions, which is a constant for every hinge function. MARS model automatically estimates α_q and knots of hinge functions for the best prediction performance in the case. If the hinge function is a constant for one input, it means that this input is not used in the model. As a result, there is one term (the constant) for the input if it is not used and there are two terms ($\alpha_{q1}\max(0, \mathrm{x}-\mathrm{C})$, $\alpha_{q2}\max(0, \mathrm{C}-\mathrm{x})$) for the input if it is used in a simple MARS model. The number of terms in MARS model as well as its α_q and knots will be determined when training the model. MARS model performs better than linear regression models when there are strong non-linearities between inputs and outputs. However, it can also be negatively affected by outliers since it is sensitive to them. After deleting outliers, MARS model will not perform worse than linear regression models theoretically because it can also deal with redundant inputs with the constant term, which is suitable for both linear and non-linear problems.

In this case, 21 terms and 10 inputs are tuned (RLA0, RLA4, RLA5 are not used) with the highest prediction accuracy of MARS model to be 0.85 and 0.83 for DTC and DTS prediction respectively, in terms of R2. Although three inputs (RLA0, RLA4, RLA5) of high correlation with other inputs (RLA1, RLA2, RLA3) are not used in MARS model, it still performs better than four linear regression models.

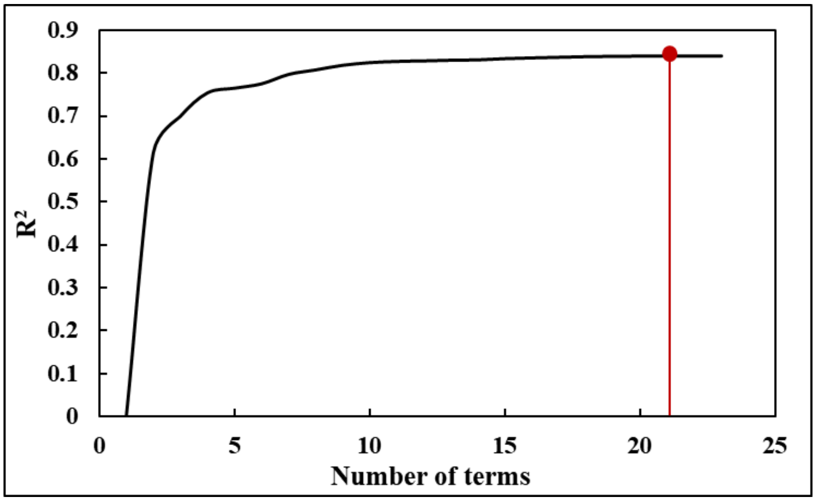


Figure 5.5 Tuning the parameter of number of terms in the MARS model.

5.2.6 Artificial Neural Network (ANN) Model

ANN model is a widely-used machine learning model with various structures, algorithms and applications. In the ANN model, the first layer is input layer and the last layer is output layer. The number of hidden layers in the middle and the number of neurons in each hidden layer should be decided according to the number of inputs, outputs and the complexity of the problem. An arithmetic sequence of the number of neurons in each hidden layer generates high prediction accuracy (Demuth at al. 2014). In my case, there are 13 inputs and 2 outputs for sonic log prediction, two hidden layers is set in my model, which is enough for function approximation problems, so that there are 9 neurons in the first hidden layer and 5 neurons in the second hidden layer. The ANN model with 3 hidden layers does not outperform the one with 2 hidden layers (Figure 5.6).

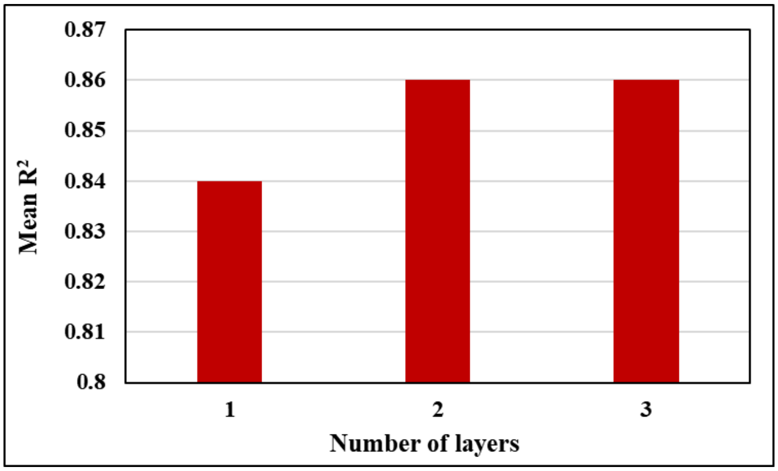


Figure 5.6 Comparison of prediction performances of ANN models with different number of hidden layers.

Several algorithms can be utilized as the training function to adjust weights and biases of the neurons for minimizing certain loss functions of the ANN model. Best performing ANN model requires the minimization of the loss functions using the training functions. Levenberg-Marquardt (LM) backpropagation (Chamjangali et al., 2007) and

Conjugate Gradient (CG) backpropagation (Cheng et al., 2005) are the two most widely used training functions. CG backpropagation is applied as the training function after comparison of training time with two algorithms. Training time of CG backpropagation is about the half compared with that of LM backpropagation. To be specific, scaled CG backpropagation (Cheng et al., 2005) is selected and tested as the best training function for the ANN models implemented in this thesis.

Overfitting is the main problem when minimizing the value of loss function during iterations. ANN model cannot recognize the appropriate relationships between inputs and outputs because of overfitting. The following loss function is designed to avoid overfitting, which is expressed as

$$L = \sum_{i=1}^{n} (y_{i,j} - \hat{y}_{i,j})^2 + \lambda \sum_{j=1}^{o} \sigma_j^2 \quad \text{(5.7)}$$

where n is the number of depths in a well, j = 1 and 2 represents DTC and DTS, respectively, o = 2 is the total number of output logs, λ is the penalty parameter, $y_{i,j}$ is the sonic log j measurement at depth i, $\hat{y}_{i,j}$ is the of output log j prediction at depth i, and σ_j^2 is the variance of the predicted output log j, such that

$$\sigma_j^2 = \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_{i,j} - \mu_j)^2 \quad \text{(5.8)}$$

where μ_j is the mean of the predicted output log j. Regularization method is utilized to introduce a penalty parameter in the loss function to avoid overfitting and ensure a balanced bias-variance tradeoff. λ ranges from 0 to 1, which is set based on extensive numerical experiments. After tuning λ, 0.01 is selected, which is a very small number because of simplicity of the model.

In this case, the prediction performance of ANN model is 0.87 and 0.85 for DTC and DTS prediction respectively, in terms of R2. Although it is more accurate than MARS model, ANN model takes more computational time as well.

5.3 Case Study

5.3.1 Prediction Results of Six Models

The comparison of prediction performance of six models are shown in Table 5.1 and visualized in Figure 5.7. In Well 1, ANN model performs the best for predicting DTC and DTS logs with R^2 of 0.87 and 0.85, respectively (Figure 5.8). MARS performs second among all models, followed by OLS and PLS. LASSO and ElasticNet perform the worst. Six models are then deployed in the 1460-feet depth interval of Well 2 in the same reservoir. The prediction accuracy ranking does not change much. But PLS performs worse than OLS and ElasticNet performs worse than LASSO. Every model has a different accuracy drop when trained in Well 1 and deployed in Well 2 but there accuracy ranking keep almost the same, for example, ANN performs the best both in Well 1 and Well 2. The accuracy drop of PLS is much larger than that of others, especially LASSO and ElasticNet, which show that the combination of original inputs may not be a stable method to decrease dimension of inputs.

Computational time of six models is also compared for the balance of the simplicity of models and their accuracy (Table 5.2). ANN cost the most in computational time. PLS is the second because the combination of inputs cost a lot of time. The rest models all cost little time in training.

Table 5.1 Prediction performance of six models trained and tested in Well 1 and applied in Well 2.

Accuracy		OLS	PLS	LASSO	ElasticNet	MARS	ANN
Well 1	DTC	0.830	0.830	0.791	0.791	0.847	0.870
	DTS	0.803	0.803	0.756	0.753	0.831	0.848
Well 2	DTC	0.804	0.790	0.778	0.774	0.816	0.850
	DTS	0.794	0.769	0.763	0.755	0.806	0.840

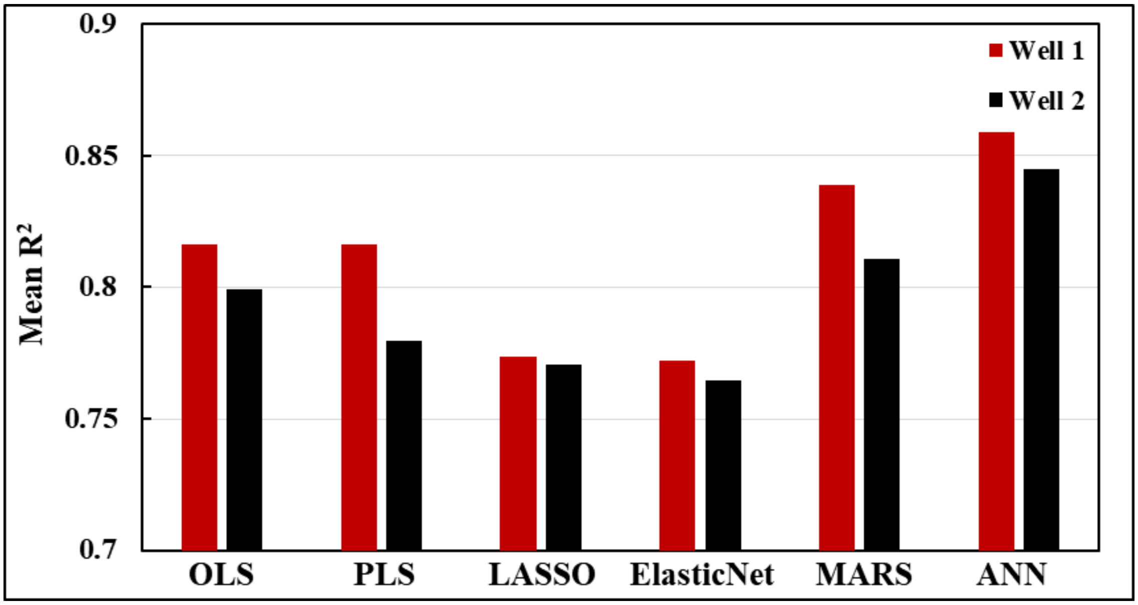


Figure 5.7 The comparison of prediction performance of six models trained and tested in Well 1 and deployed in Well 2.

Table 5.2 Comparison of computational time for training the six models in Well 1.

	OLS	PLS	LASSO	ElasticNet	MARS	ANN
Computational time (s)	0.09	3.57	0.03	0.04	0.14	4.17

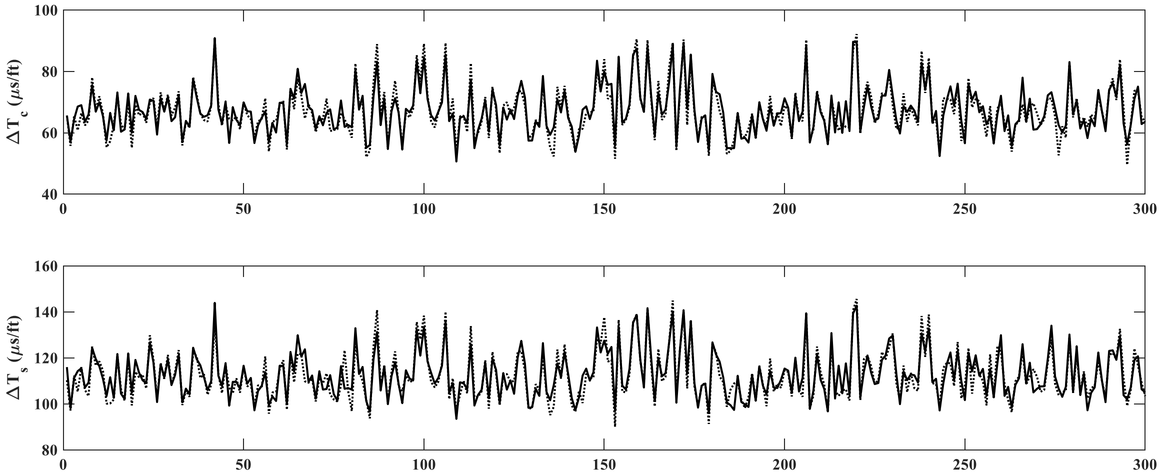


Figure 5.8 Comparison of original (dashed) and predicted (solid) DTC and DTS logs in Well 1, when an ANN model is trained and tested in Well 1 to generate the DTC and DTS logs.

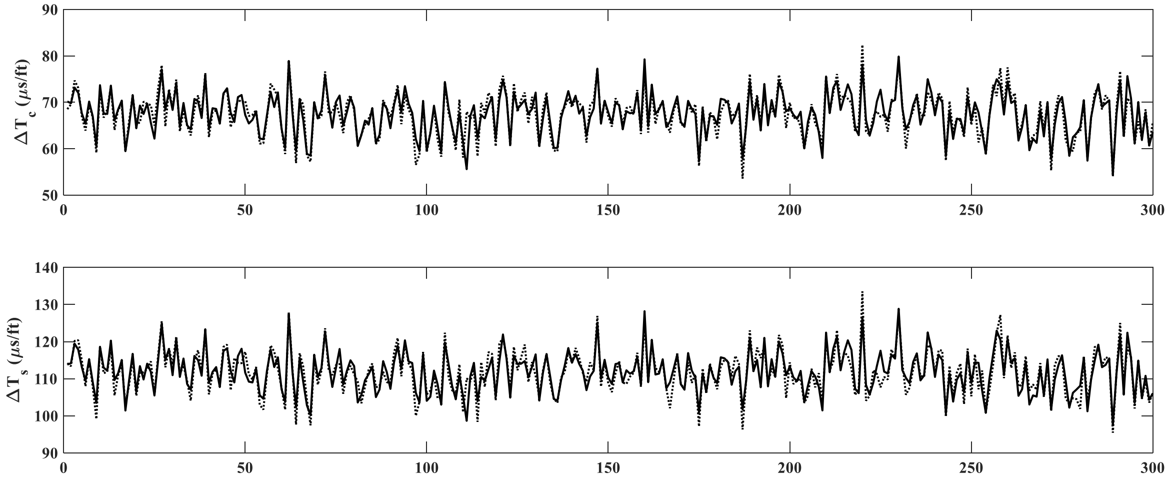


Figure 5.9 Comparison of original (dashed) and predicted (solid) DTC and DTS logs in Well 2, when an ANN model is trained and tested in Well 1 and deployed in Well 2 to generate the DTC and DTS logs.

5.3.2 Analysis of the Results

ANN model performs the best among six models, following by MARS model the second, OLS and PLS models with medium accuracy, and LASSO and ElasticNet models with low accuracy. There are several reasons for the prediction performance of six models being like that:

The complex structure of the ANN model with its multiple neurons and layers make it the best performing model among the six models.

MARS model performs better than the other four regression models because MARS model take non-linearities into consideration while the other four models are all linear regression models.

OLS and PLS models have the same prediction performance because m is tuned to be equal to p in PLS model, which implies that there are no obvious redundant inputs. When the number of constructed components is the same as the number of original inputs, the prediction performance of the two models must be the same theoretically.

The accuracy of LASSO and ElasticNet models are the same because their penalty parameters are tuned to be close to each other ($\lambda \approx \lambda_1$ and $\lambda_2 \approx 0$). When $\lambda_2 = 0$, the ElasticNet model becomes the LASSO model.

OLS and PLS models outperform LASSO and ElasticNet models because the former two utilize all inputs to build models while the latter two penalize inputs by different parameters. There are no obvious redundant inputs so that penalty is not necessary.

The prediction performances of six models for DTC are always better than those for DTS. One possible reason is that the higher speed of compressional wave leads to cleaner acquisition of DTC without the interference of shear waves. Another possible reason is that DTC is less affected by cracks, unconsolidated formations and borehole fluids than DTS, which makes DTC more reliable with less noise (Cheng et al., 1992).

The structure of models is more important than algorithms for good prediction performance. For example, both non-linear models perform better than four linear models although similar loss functions are applied in them.

5.3.3 Error Distribution Comparison of Prediction Performance for Six Models

Relative Error (RE) is used to evaluate the prediction performance as well. RE is formulated as

$$RE = \frac{|P-M|}{M} \qquad (5.9)$$

where P is the predicted value and M is the measured value. Mean RE for DTC and DTS are calculated at all depths, resulting in one RE at each depth. When mean RE is less than 0.1, the depth belongs to the category of good prediction performance. When mean RE is at the range of [0.1, 0.2], the depth belongs to the category of medium prediction performance. When mean RE is larger than 0.2, the depth belongs to the category of poor prediction performance. Poor prediction performance happens in every model from around 1250 to 1800 ft below the top of the formation depth under investigation. Medium prediction performance happens in every model from around 3800 to 4240 ft in depth. In conclusion, all models perform similarly at same depths. At the depths corresponding to 1250 to 1800 ft in depth, there should be noise either in inputs or outputs.

Three statistical measurements are used to analyze the relationship between statistical properties of logs and prediction performance. The 4240-feet depth interval in Well 1 is divided into four big parts (I: 0-1250 ft; II: 1250-1800 ft; III: 1800-3800 ft; IV:

3800-4240 ft below the top of the formation depth under investigation). From prediction performance of six models in Figure 5.10, part I and III belong to good prediction performance, part II belongs to poor prediction performance and part IV belongs to medium prediction performance. Three statistical measurements are selected for analysis including mean (μ), coefficient of variation (S_d/μ) and the absolute value of skewness (|s|). S_d/μ is standard deviation over mean value. In this case, larger μ of GR, porosity, DTC, DTS and smaller μ of bulk density tend to result in poorer prediction performance. S_d/μ and skewness do not affect prediction performance much. High absolute value of skewness of porosity tends to result in poorer prediction performance.

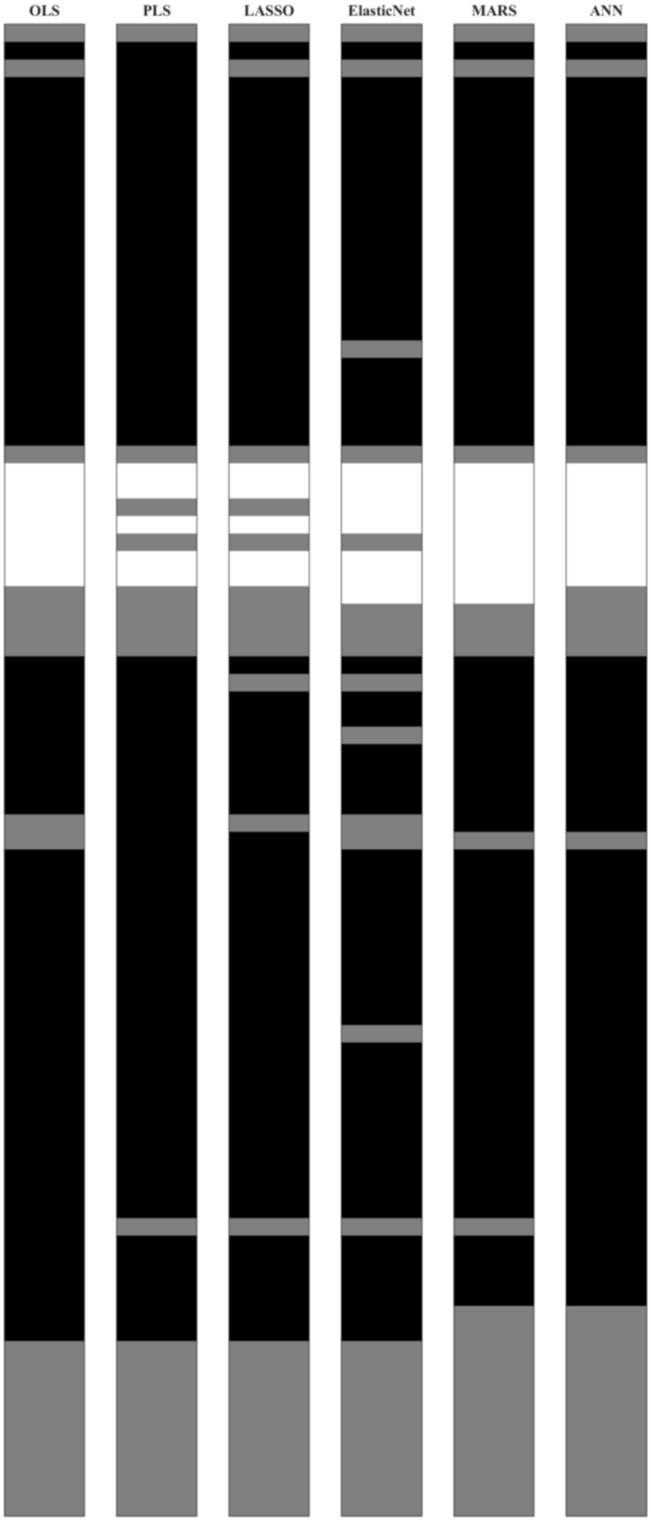


Figure 5.10 Comparison of RE distribution in six models all through the 4240-feet depth interval in Well 1, black represents good prediction performance depths, gray represents medium prediction performance depths, white represents poor prediction performance depths.

Table 5.3 Statistical description of all inputs and outputs (I: 0-1250 ft; II: 1250-1800 ft; III: 1800-3800 ft; IV: 3800-4240 ft).

	μ				S_d/μ				$\lvert s\rvert$			
	I	II	III	IV	I	II	III	IV	I	II	III	IV
GR	68.217	95.620	70.666	83.113	0.430	0.361	0.364	0.259	1.430	0.402	0.616	0.779
DCAL	1.909	0.786	0.765	0.120	0.543	0.132	0.412	1.565	1.511	0.166	1.332	0.222
DPHZ	0.089	0.141	0.064	0.082	0.468	0.353	0.550	0.347	0.183	0.533	0.191	1.214
NPOR	0.095	0.177	0.092	0.116	0.427	0.410	0.513	0.312	0.237	0.403	0.166	0.396
PEFZ	3.194	3.303	3.467	2.923	0.305	0.184	0.243	0.147	1.668	1.043	1.278	1.763
RHOZ	2.566	2.483	2.607	2.578	0.026	0.032	0.022	0.018	0.183	0.533	0.191	1.213
RLA0	0.031	0.035	0.033	0.033	0.120	0.024	0.059	0.033	0.075	0.667	0.481	0.049
RLA1	11.670	15.717	23.007	16.533	0.618	0.412	0.476	1.101	0.703	0.754	0.168	2.279
RLA2	33.677	47.158	86.168	21.517	1.369	0.940	0.932	1.714	2.378	1.310	1.865	5.541
RLA3	61.603	82.055	143.065	23.881	1.825	1.222	1.148	1.967	3.193	1.956	1.929	6.489
RLA4	134.719	117.122	228.613	24.890	2.781	1.555	1.828	2.190	5.710	3.326	4.705	7.641
RLA5	187.924	129.367	247.366	24.707	2.894	1.472	2.084	2.108	6.013	3.541	6.018	7.229
DTC	64.040	78.444	64.504	71.172	0.080	0.132	0.103	0.064	0.065	0.753	0.187	0.572
DTS	109.783	126.857	108.676	118.544	0.080	0.100	0.076	0.058	1.251	0.419	0.191	0.423

5.4 Sensitivity Study

5.4.1 Noise of Inputs and Outputs

Every input log except the flag is added with 10% Gaussian noise separately. The flag obtained from interval depths is not included because they are classified into 13 categories with 13 integers. All data of the flag will be meaningless when 10% noise is added to them because they will be real numbers instead of integers. 10% means that the standard deviation of the noise distribution, which is Gaussian distribution, is equal to 10% of the input value itself. For every single value of every input, the noise distribution is not the same and one value of noise is selected randomly from the Gaussian distribution. In addition, the noise of every input value is different every time the noise the randomly selected from the noise distribution. As a result, parallel computing is used in this case to set random noise on logs for 50 times and average the result. The effect of noise of every input log on prediction accuracy is recorded and their ranking sequence is shown in Figure 5.11. Generally, GR is affected the most by noise and the rest are similarly affected by noise. This does not mean that GR is the most important and the rest are of similar importance. For example, RLA0-5 are of high correlation between each other and one of them with noise will not affect the prediction performance much when others exist. So are DPHZ and NPOR. The accuracy drop comparison show that all input logs are important to prediction performance and they are all sensitive to noise.

DTC and DTS logs are added with 10% and 20% Gaussian noise separately as well. When DTC is added with noise, the prediction accuracy of DTS will not drop, and vice versa. 10% noise is added to DTC and DTS logs and accuracy is dropped by 41.4% and 52.8% respectively. 20% noise is then added to DTC and DTS logs as well and accuracy is dropped by 73.3% and 81.0% respectively. In conclusion, outputs are much more sensitive to noise than inputs. DTS is more sensitive to noise than DTC.

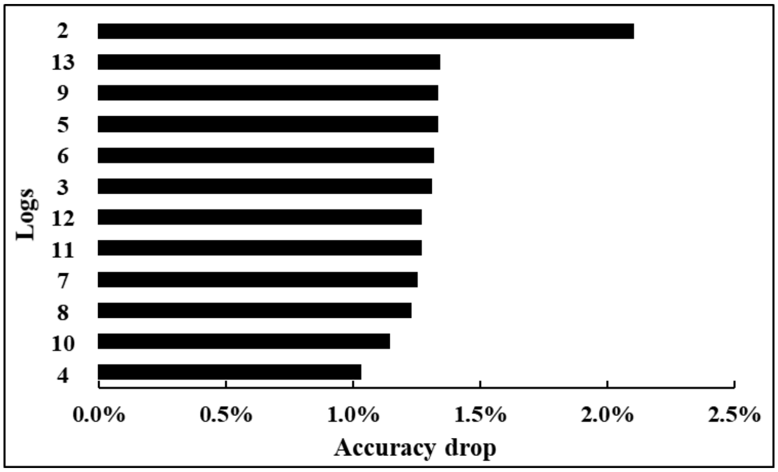


Figure 5.11 Comparison of reduction in prediction performance when 10% Gaussian noise is added to inputs one at a time in the ANN model. Each integer on y-axis identifies the removal of a specific log (2: GR; 3: DCAL; 4: DPHZ; 5: NPOR; 6: PEFZ; 7: RHOZ; 8: RLA0; 9: RLA1; 10: RLA2; 11: RLA3; 12: RLA4; 13: RLA5).

5.4.2 Comparison of Models with Noisy Inputs

10% Gaussian noise is added to all inputs together before training and testing six shallow learning models. Accuracy drop of OLS, LASSO and ElasticNet models are lower than that of PLS, MARS and ANN models. The former three models (OLS, LASSO and ElasticNet) are less sensitive to noise because they are all simple linear regressions with simple model structures. In LASSO and ElasticNet models, penalty parameters are introduced into loss functions without change of model structures. As a result, when noise is added to inputs, weights and biases of these models will not change much (around 3% accuracy drop in terms of R^2). The latter three models (PLS, MARS and ANN) are, however, more sensitive to noise. PLS model constructs a set of linear combinations of original inputs in order to reduce dimension, which will be negatively affected by noise of inputs. MARS and ANN models both consider non-linearities during the training step. Noise of inputs will be more detected by them, treating noise as possible relationship between inputs and outputs, which is overfitting. In conclusion, simple linear regression models are more robust against noise while complicated models tend to be more sensitive to noise.

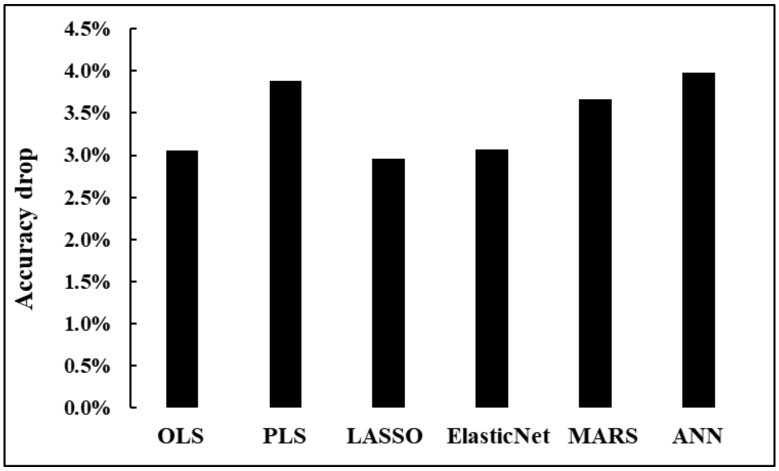


Figure 5.12 Comparison of reduction in prediction performance when 10% Gaussian noise is added to all inputs in six models.

5.4.3 Comparison of Models with Fewer Data

Three different sizes of dataset are used to train and test the six models separately. 8481 data are the entire dataset in Well 1. About half (4000 data) of them are randomly selected from the entire dataset as the second dataset. One quarter (2000 data) dataset are randomly selected from the second one as the third one to train six models. Accuracy in terms of R^2 drops each time when the size of dataset decreases in six models. OLS, LASSO and ElasticNet models are less sensitive to dataset size while PLS, MARS and ANN models will be more affected by it. 2000 data is not a small size in a linear regression problem with 13 inputs and 2 outputs. As a result, the prediction performance of the former three mentioned models will not change much when dataset enlarges more than that. However, the prediction performance of the latter three mentioned models will improve apparently when dataset enlarges, in which more data are needed for robust non-linear models or the linear model with construction of new inputs. It is similar to their performance to noise of inputs. models which are more robust to noise tend to be more robust to dataset size, and vice versa.

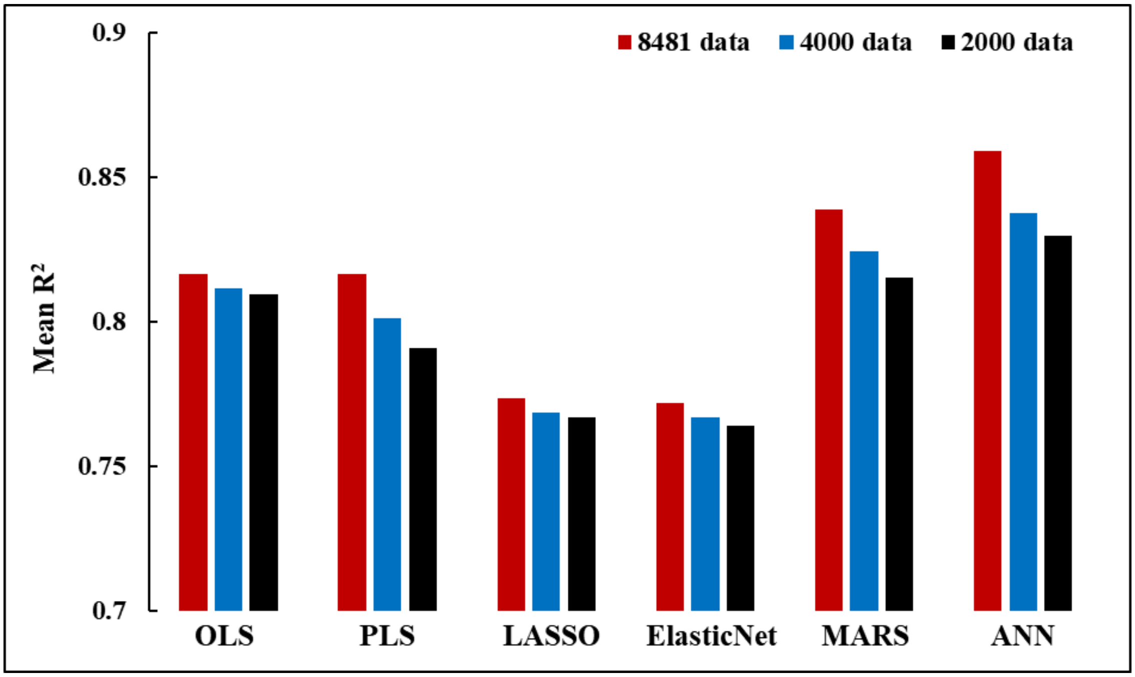


Figure 5.13 Comparison of reduction in prediction performance when different sizes of dataset are used in six models.

5.5 Conclusions

Six shallowing learning models are selected to predict DTC and DTS logs at the same time processing 13 conventional and easy-to-acquire logs from the well in PB. OLS, PLS, LASSO and ElasticNet models are four linear regression models while MARS and ANN can also deal with problems of high non-linearities. It is the first time that a data-driven method for purposes of shale reservoir geomechanical characterization was proposed based on an extensive comparison of the prediction performances of six shallow learning models. After comparison, ANN model performs the best for predicting DTC and DTS logs, with R^2 of 0.87 and 0.85 in Well 1 and with R^2 of 0.85 and 0.84 in Well 2, respectively. In this case, larger μ of GR, porosity, DTC, DTS, smaller μ of bulk density and larger absolute value of skewness of porosity tend to result in poorer prediction performance. Prediction performance for DTC is always better than that for DTS because of higher speed and more reliability. Simple linear regression models tend to be more robust to both noise and dataset size while complicated models tend to be more sensitive to them.

Chapter 6: Conclusions, Limitations and Recommendations

The objective of this research is to apply machine learning methods to synthetically generate certain well logs, such as NMR, dielectric dispersion and sonic logs.

In the first part, 12 conventional logs, 10 inversion-derived logs, and 5 qualitative log-derived flags were processed by two distinct neural network models to generate the NMR T_2 distribution responses, which approximate the in-situ fluid-filled pore size distribution in hydrocarbon-bearing BPS. The first predictive model generates T_2 distribution discretized into 64 bin amplitudes, whereas the second predictive model generates the 6 fitting parameters that characterize the T_2 distribution as a sum of two Gaussian distributions. The first predictive model performs more accurately, exhibiting median R^2 of 0.8549 during testing, compared to the second one, exhibiting median R^2 of 0.7584. However, the second model has lower computational cost compared to the first model. Input data used in my predictive models were limited in quantity and prone to noise and uncertainty due to the subsurface borehole conditions in which they were acquired. Nonetheless, the two predictive models exhibit good prediction performance. A few reservoir properties, ϕ_N, $T_{2,gm}$ and $T_{2,gm}\phi_N^2$, were derived from the synthetic T_2 distribution at reasonable accuracies. Complicated pore size distribution caused by complex grain size distribution and textures can impede good prediction performance of NMR T_2 distribution. Different thickness of intervals is also one reason for different prediction accuracy. This study provides a first-of-its-kind technique to generate in-situ fluid-filled pore size distribution, approximated as NMR T_2 distribution, in hydrocarbon-bearing shale reservoirs using neural network models that process conventional easy-to-acquire logs and inversion-derived mineral volume fractions and fluid saturations. The proposed method holds value in the absence of NMR logging tool due to financial and operational challenges.

In the second part, three neural-network-based predictive methods are developed to process 15 conventional logs and generate 8 synthetic dielectric dispersion (DD) logs, comprising 4 conductive dispersion and 4 permittivity dispersion logs. The first method simultaneously predicts the 8 DD logs. The second method first processes the 15 conventional logs to predict the 4 conductivity dispersion logs, which are then processed along with the 15 conventional logs to predict the 4 permittivity dispersion logs. In comparison to the first method, prediction performance of the second method is 8.5% worse for conductivity dispersion and 6.2% better for permittivity dispersion. The third method exhibits the best predictive performance for the generation of the 8 synthetic DD logs. In comparison to the first method, prediction performance of the third method is 0.8% better for conductivity dispersion and 8.5% better for permittivity dispersion. Performances of these models are adversely affected by the noise in the logs and the limited amount of data available for the training purposes. High water salinity in formations impedes good predictive performance. Inclusion of NMR logging data does not improve the predictive performance. NRMSE is a better indicator of predictive performance compared to R^2 in this case. Low resistivity, high porosity, high relative dielectric permittivity, large dielectric dispersion, and low skewness and large coefficient of variation of conventional log inputs facilitate high prediction performance for the 8 synthetic dielectric dispersion logs. Training the third predictive method in one well and then deploying it in another well for generating the 8 DD logs is feasible, such that the

NRMSE of conductivity dispersion logs drops by 6% and that of permittivity dispersion logs drops by 4.5% compared to the baseline. Resistivity logs of various depths of investigation, neutron porosity log, and compressional and shear travel time logs are the most important inputs for the DD log generation. Deep resistivity logs are more important than the shallow ones for the prediction. Noise in resistivity, gamma ray, shear travel time, and dielectric dispersion logs adversely influences the prediction performance.

In the third part, six shallowing learning models are selected to predict DTC and DTS at the same time processing 13 conventional and easy-to-acquire logs from the well in a shale reservoir in PB. OLS, PLS, LASSO and ElasticNet models are four linear regression models while MARS and ANN can also deal with problems of high non-linearities. It is the first time that a data-driven method for purposes of shale reservoir geomechanical characterization was proposed based on an extensive comparison of the prediction performances of six shallow learning models. After comparison, ANN model performs the best both in Well 1 and Well 2. Some conclusions can be made after result analysis and sensitivity analysis: (1) in this case, larger μ of GR, porosity, DTC, DTS, smaller μ of bulk density and larger absolute value of skewness of porosity tend to result in poorer prediction performance; (2) the structure of models is more important than algorithms for good prediction performance; (3) prediction performance for DTC is always better than that for DTS because of higher speed and more reliability; (4) simple linear regression models tend to be more robust to both noise and dataset size while complicated models tend to be more sensitive to them. This study will enable people to obtain improved geomechanical characterization under data constraints by developing and identifying the best performing machine-learning model for the prediction of DTC and DTS logs when sonic logging tool is not available due to operational and financial challenges.

However, there are also some limitations and recommendations for future work:

(1) Inputs are selected when they are available and easy to be acquired. The most related inputs are selected according to physical properties. The most related inputs can be selected with the combination of other algorithms and statistical methods for prediction of specific well logs.

(2) The number of wells and the size of data are limited in this thesis. If more wells in a reservoir are available, models can be trained and tested in one well and deployed in other wells for verification.

(3) Limited analyzation is done for the importance of inputs as well as the effect of noise. Further research can be done to analyze the importance of inputs to prediction performance and effect of noise on prediction performance for different well logs as well as their theoretical supporting reasons.

(4) Limited analyzation is done for the statistical characteristics of both inputs and outputs. Further research can be done to analyze the relationship between statistical characteristics of both inputs and outputs and prediction performance of well logs.

(5) Only different ANN models are compared in the first and second research. More models can be built and compared together for predicting NMR T_2 distribution and DD logs. The structures and algorithms of models should be analyzed to find theoretical supporting reasons to select the best model.

References

Al-Bulushi, N., Araujo, M., Kraaijveld, M., & Jing, X. D., 2007. Predicting water saturation using artificial neural networks (ANNS). In SPWLA Middle East Regional Symposium. Society of Petrophysicists and Well-Log Analysts.

Alegre, L., 1991. Potential applications for artificial intelligence in the petroleum industry. Journal of Petroleum Technology 43, no. 11: 1-306.

Asoodeh M., & Bagheripour P., 2012. Prediction of compressional, shear, and stoneley wave velocities from conventional well log data using a committee machine with intelligent systems. Rock mechanics and rock engineering, 45(1), 45-63.

Asquith, G. B., Krygowski, D., & Gibson, C. R., 2004. Basic well log analysis (Vol. 16). Tulsa: American Association of Petroleum Geologists.

Baines V., Bootle R., Pritchard T., Macintyre H., & Lovell M. A., 2008. Predicting Shear And Compressional Velocities In Thin Beds. In 49th Annual Logging Symposium. Society of Petrophysicists and Well-Log Analysts.

Bhatt, A., & Helle, H. B., 2002. Committee neural networks for porosity and permeability prediction from well logs. Geophysical Prospecting, 50(6), 645-660.

Brovko A V, Murphy E K, Yakovlev V V., 2009, Waveguide microwave imaging: Neural network reconstruction of functional 2-D permittivity profiles[J]. IEEE Transactions on Microwave Theory and Techniques, 57(2): 406-414.

Chamjangali, M. A., Beglari, M., & Bagherian, G., 2007. Prediction of cytotoxicity data (CC 50) of anti-HIV 5-pheny-l-phenylamino-1H-imidazole derivatives by artificial neural network trained with Levenberg–Marquardt algorithm. Journal of Molecular Graphics and Modelling, 26(1), 360-367.

Chen Q, Huang K M, Yang X, et al., 2011, An artificial nerve network realization in the measurement of material permittivity. Progress In Electromagnetics Research, 116: 347-361.

Cheng, C., Chau, K., Sun, Y., & Lin, J., 2005. Long-term prediction of discharges in Manwan Reservoir using artificial neural network models. Advances in neural networks–ISNN 2005, 975-975.

Cheng C. H., Paillet F. L., & Pennington W. D., 1992. Acoustic-waveform Logging-advances in Theory And Application. The Log Analyst, 33(03).

Demuth H B, Beale M H, De Jess O, et al., 2014. Neural network design[M]. Martin Hagan.

Di, J., 2015. Permeability characterization and prediction in a tight oil reservoir, Edson Field, Alberta (Doctoral dissertation, University of Calgary).

Draper N. R., & Smith, H., 2014, Applied regression analysis(Vol. 326). John Wiley & Sons.

Elshafei, M., & Hamada, G. M., 2009. Petrophysical Properties Determination of Tight Gas Sands From NMR Data Using Artificial Neural Network. In SPE Western Regional Meeting. Society of Petroleum Engineers.

Ganapathi A, Chen Y, Fox A, et al., 2010, Statistics-driven workload modeling for the cloud. Data Engineering Workshops (ICDEW), 2010 IEEE 26th International Conference, 87-92.

Genty, C., 2006. Distinguishing carbonate reservoir pore facies with nuclear magnetic resonance as an aid to identify candidates for acid stimulation (Doctoral dissertation, Texas A&M University).

Genty, C., Jensen, J. L., & Ahr, W. M., 2007. Distinguishing carbonate reservoir pore facies with nuclear magnetic resonance measurements. Natural Resources Research, 16(1), 45-54.

Golsanami, N., Kadkhodaie-Ilkhchi, A., Sharghi, Y., & Zeinali, M., 2014. Estimating NMR T 2 distribution data from well log data with the use of a committee machine approach: A case study from the Asmari formation in the Zagros Basin, Iran. Journal of Petroleum Science and Engineering, 114, 38-51.

Han, Y., Misra, S., & Simpson, G., 2017, Dielectric Dispersion Log Interpretation in Bakken Petroleum System. In SPWLA 58th Annual Logging Symposium. Society of Petrophysicists and Well-Log Analysts.

Han, Y., & Misra, S., 2018, Joint petrophysical inversion of multifrequency conductivity and permittivity logs derived from subsurface galvanic, induction, propagation, and dielectric dispersion measurements. Geophysics, 83(3), 1-63.

Hasan A, Peterson A F., 2011, Measurement of complex permittivity using artificial neural networks. IEEE Antennas and Propagation Magazine, 53(1): 200-203.

Hizem M, Budan H, Deville B, et al., 2008, Dielectric dispersion: A new wireline petrophysical measurement. SPE Annual Technical Conference and Exhibition. Society of Petroleum Engineers.

Iverson W. P., & Walker J. N., 1992. Shear and compressional logs derived from nuclear logs. The Log Analyst, 33(01).

Keys R. G., & Xu S., 2002. An approximation for the Xu-White velocity model. Geophysics, 67(5), 1406-1414.

Kuhn M, Johnson K., 2013. Applied predictive modeling[M]. New York: Springer.

Labani, M. M., Kadkhodaie-Ilkhchi, A., & Salahshoor, K., 2010. Estimation of NMR log parameters from conventional well log data using a committee machine with intelligent systems: a case study from the Iranian part of the South Pars gas field, Persian Gulf Basin. Journal of Petroleum Science and Engineering, 72(1), 175-185.

Li, H., & Misra, S., 2017a, Prediction of subsurface NMR T2 distribution from formation-mineral composition using variational autoencoder. In 2017 SEG International Exposition and Annual Meeting. Society of Exploration Geophysicists.

Li, H., & Misra, S., 2017b, Prediction of subsurface NMR T2 distributions in a shale petroleum system using variational autoencoder-based neural networks. IEEE Geoscience and Remote Sensing Letters, 14(12), 2395-2397.

Mahmoud, A. A. A., Elkatatny, S., Mahmoud, M., Omar, M., Abdulraheem, A., & Ali, A., 2017. Determination of the total organic carbon (TOC) based on conventional well logs using artificial neural network. International Journal of Coal Geology.

Maleki S., Moradzadeh A., Riabi R. G., Gholami R., & Sadeghzadeh F., 2014. Prediction of shear wave velocity using empirical correlations and artificial intelligence methods. NRIAG Journal of Astronomy and Geophysics, 3(1), 70-81.

McNeal R P., 1965, Hydrodynamics of the Permian basin.

Mendenhall, W. M., & Sincich, T. L., 2016. Statistics for Engineering and the Sciences. Chapman and Hall/CRC.

Mohaghegh, S., Richardson, M., & Ameri, S., 2001. Use of intelligent systems in reservoir characterization via synthetic magnetic resonance logs. Journal of Petroleum Science and Engineering, 29(3), 189-204.

Nitta T., 2009, Complex-valued neural networks. IGI Global.

Pirie, I., Horkowitz, J., Simpson, G., & Hohman, J., 2016. Advanced methods for the evaluation of a hybrid-type unconventional play: The Bakken petroleum system. Interpretation, 4(2), SF93-SF111.

Rezaee M. R., Ilkhchi A. K., & Barabadi A., 2007. Prediction of shear wave velocity from petrophysical data utilizing intelligent systems: An example from a sandstone reservoir of Carnarvon Basin, Australia. Journal of Petroleum Science and Engineering, 55(3-4), 201-212.

Russell, S. J., & Norvig, P., 2016. Artificial intelligence: a modern approach. Malaysia; Pearson Education Limited.

Salazar, J. P., & Romero, P. A., 2001. NMR measurements in carbonates samples and core-logs correlations using artificial neural nets. In SPE Annual Technical Conference and Exhibition. Society of Petroleum Engineers.

Simpson, G., Hohman, J., Pirie, I., & Horkowitz, J., 2015. Using advanced logging measurements to develop a robust petrophysical model for the Bakken Petroleum System. In SPWLA 56th Annual Logging Symposium. Society of Petrophysicists and Well-Log Analysts.

Simpson G, Hohman J, Pirie I, et al., 2015, Using advanced logging measurements to develop a robust petrophysical model for the Bakken Petroleum System. SPWLA 56th Annual Logging Symposium. Society of Petrophysicists and Well-Log Analysts.

Tathed, P., Han, Y., & Misra, S., 2018, Hydrocarbon saturation in upper Wolfcamp shale formation. Fuel, 219, 375-388.

Vapnik V., 2013, The nature of statistical learning theory. Springer science & business media.

Willis M. E., & Toksoez M. N., 1983. Automatic P and S velocity determination from full waveform digital acoustic logs. Geophysics, 48(12), 1631-1644.

Xu S., & White R. E., 1995. A new velocity model for clay-sand mixtures. Geophysical prospecting, 43(1), 91-118.

Yang C C, 2005, Bose N K. Landmine detection and classification with complex-valued hybrid neural network using scattering parameters dataset. IEEE transactions on neural networks, 16(3): 743-753.

Appendix

ANN: Artificial Neural Network
AT10: Induction Resistivity Logs at 10-inch
AT90: Induction Resistivity Logs at 90-inch
BP: Back Propagation
BPS: Bakken Petroleum System
BS: Bakken Shale
CG: Conjugate Gradient algorithm
CMIS: Committee Machine with Intelligent Systems
CVNN: Complex-Valued Neural Network
DCAL: Caliper Log
DD: Dielectric Dispersion
DPHZ: Density Porosity Log
DTC: Delta-T Compressional Log
DTS: Delta-T Shear Log
EM: Electromagnetic (waves)
GR: Gamma Ray Log
KNN: K-Neareat Neighbor algorithm
LASSO: Least Absolute Shrinkage and Selection Operator
LBS: Lower Bakken Shale
LM: Levenberg-Marquardt algorithm
MARS: Multivariate Adaptive Regression Splines
MBS: Middle Bakken Shale
NMR: Nuclear Magnetic Resonance
NPOR: Neutron Porosity Log
NRMSE: Normalized Root Mean Square Error
OLS: Ordinary Least Squares
PB: Permian Basin
PEFZ: Photoelectric Factor Log
PLS: Partial Least Squares
RE: Relative Error
RHOZ: Bulk Density Log
RLA0-5: Laterolog resistivity at different depths of investigation
SSE: Sum of Squared Errors
TF: Three Forks
TOC: Total Organic Carbon
UBS: Upper Bakken Shale
VCL: Volume of Clay Log

VPVS: Shear to Compressional Velocity Ratio