โครงการการเรียนการสอนเพื่อเสริมประสบการณ์

การจำลองข้อมูลการหยั่งธรณีหลุมเจาะสังเคราะห์จากรัฐแคนซัส ประเทศสหรัฐอเมริกา โดยวิธีการเรียนรู้ของเครื่อง

โดย

นางสาวกมลพรรณ พันธุ์นิธิประเสริฐ

เลขประจำตัวนิสิต 5932701823

โครงการนี้เป็นส่วนหนึ่งของการศึกษาระดับปริญญาตรี
ภาควิชาธรณีวิทยา คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2562

การจำลองข้อมูลการหยั่งธรณีหลุมเจาะสังเคราะห์จากรัฐแคนซัส ประเทศสหรัฐอเมริกา โดยวิธีการเรียนรู้ของเครื่อง

นางสาวกมลพรรณ พันธุ์นิธิประเสริฐ

โครงงานนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตร์บัณฑิต
ภาควิชาธรณีวิทยา คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2562

RECONSTRUCTION OF SYNTHETIC WELL LOGS FROM KANSAS, USA USING MACHINE LEARNING APPROACH

MISS KAMONPHAN PHANNITHIPRASERT

A Project Submitted in Partial Fulfilment of the Requirements

for the Degree of Bachelor of Science Program in Geology

Department of Geology, Faculty of Science, Chulalongkorn University

Academic Year 2019

หัวข้อโครงงาน	การจำลองข้อมูลการหยั่งธรณีหลุมเจาะสังเคราะห์จากรัฐแคนซัส ประเทศสหรัฐอเมริกา โดยวิธีการเรียนรู้ของเครื่อง
โดย	นางสาวกมลพรรณ พันธุ์นิธิประเสริฐ
สาขาวิชา	ธรณีวิทยา
อาจารย์ที่ปรึกษาโครงงาน	รองศาสตราจารย์ ดร.สันติ ภัยหลบลี้
อาจารย์ที่ปรึกษาโครงการร่วม	ผู้ช่วยศาสตราจารย์ ดร.วรัญทร คณิตปัญญาเจริญ

วันที่ส่ง.....

วันที่อนุมัติ.....

.....

อาจารย์ที่ปรึกษาโครงงาน

(รองศาสตราจารย์ ดร.สันติ ภัยหลบลี้)

Project Title	RECONSTRUCTION OF SYNTHETIC WELL LOGS FROM KANSAS, USA USING MACHINE LEARNING APPROACH
By	Ms. Kamonphan Phannithiprasert
Field of study	Geology
Project Advisor	Associate Professor Dr. Santi Pailoplee
Co-advisor	Assistant Proferssor Dr. Waruntorn Kanitpanyacharoen

Submitted date.....

Approval date.....

.....

Project Advisor

(Associate Professor Dr. Santi Pailoplee)

กมลพรรณ พันธุ์นิธิประเสริฐ: การจำลองข้อมูลการหยั่งธรณีหลุมเจาะสังเคราะห์จากรัฐแคนซัส ประเทศสหรัฐอเมริกา โดยวิธีการเรียนรู้ของเครื่อง (RECONSTRUCTION OF SYNTHETIC WELL LOGS FROM KANSAS, USA USING MACHINE LEARNING APPROACH) อ.ที่ปรึกษาโครงงานหลัก : รองศาสตราจารย์ ดร.สันติ ภัยหลบลี้, 56 หน้า

ข้อมูลธรณีหลุมเจาะแบบโฟโตอิเล็กทริกมีความสำคัญอย่างมากในการสำรวจหาแหล่งปิโตรเลียม เนื่องจากค่าของผลบันทึกสามารภใช้เป็นตัวแทนแสดงถึงแร่หลักของชั้นหินหรือแหล่งกักเก็บได้โดยตรง เช่น ค่าผลบันทึกโฟโตอิเล็กทริกประมาณ 5 บาร์นส์ตออิเล็กตรอนสามารถเป็นตัวแทนแสดงถึงหินคาร์บอเนตในแหล่งกักเก็บ แต่อย่างไรก็ตามการหยั่งธรณีหลุมเจาะใช้งบประมาณและแรงงานปริมาณมากเพื่อเก็บข้อมูลที่จำเป็น ยิ่งไปกว่านั้น การขาดหายไปของข้อมูลที่ความลึกต่างๆก็เป็นปัญหาที่เกิดขึ้นบ่อยในระหว่างขั้นตอนการเก็บข้อมูล ดังนั้น งานวิจัยนี้จึงประยุกต์ใช้วิธีการเรียนรู้ของเครื่อง 3 วิธีดังนี้ แบบจำลองเอ็กซ์ตรีมเกรเดียนต์บูสติ้ง แบบจำลองซัพพอร์ตเวกเตอร์รีเกรสชั่นและแบบจำลองโครงข่ายประสาทเทียม เพื่อสังเคราะห์ข้อมูลธรณีหลุมเจาะแบบโฟโตอิเล็กทริกจากแอ่งอะนาร์ดาโกในรัฐแคนซัส ข้อมูลการหยั่งธรณีหลุมเจาะชนิดต่างๆ ทั้งหมด 6 ชนิดดังนี้ ข้อมูลธรณีหลุมเจาะแบบรังสีแกมมา แบบความต้านทานไฟฟ้าระดับลึก แบบศักย์ไฟฟ้าเกิดเอง แบบปริมาณรูพรุนำนวณจากความหนาแน่น แบบความหนาแน่นมวลรวมและแบบโฟโตอิเล็กทริก จำนวนมากกว่า 50,000 จุดข้อมูลจาก 12 หลุมเจาะได้ถูกนำมาใช้เพื่ออบรม ตรวจสอบและทดสอบแต่ละวิธีกรเรียนรู้ของเครื่อง ในอัตราส่วน 70:20:10 แบบจำลองโครงข่ายประสาทเทียมให้ผลลัพธ์ที่ไม่ดีนักโดยมีค่าคลาดเคลื่อนกำลังสองเฉลี่ยมากที่สุดซึ่งเท่ากับ 0.197 เนื่องจากแบบจำลองชนิดนี้ไม่สามารถจัดการกับข้อมูลที่ไม่สมดุลได้ แต่แบบจำลองเอ็กซ์ตรีมเกรเดียนต์บูสติ้งให้ผลลัพธ์ทีดีกว่าทั้งซัพพอร์ตเวกเตอร์รีเกรสชั่นและโครงข่ายประสาทเทียม โดยเอ็กซ์ตรีมเกรเดียนต์บูสติ้งมีค่าคลาดเคลื่อนกำลังสองเฉลี่ยน้อยที่สุดซึ่งเท่ากับ 0.139 และค่าสัมประสิทธิ์การตัดสินใจเท่ากับ 0.75 ผลลัพธ์นี้เกิดขึ้นได้เพราะแบบจำลองเอ็กซ์ตรีมเกรเดียนต์บูสติ้งสามารถที่จะรับมือกับข้อมูลที่ไม่สมดุลได้และยังมีความสามารถที่จะจัดลำดับความสำคัญข้อมูลธรณีหลุมเจาะชนิดต่างๆที่มีประโยชน์ในการสังเคราะห์ข้อมูลธรณีหลุมเจาะแบบโฟโตอิเล็กทริกได้ด้วยตัวเองประกอบกับมีการทำงานที่คล้ายคลึงกบการตัดสินใจของมนุษย์ ข้อมูลธรณีหลุมเจาะที่มีประโยชน์ช่วยในการสังเคราะห์ข้อมูลธรณีหลุมเจาะแบบโฟโตอิเล็กทริกมากที่สุด 3 ชนิดแรกคือ ความลึกในการเก็บข้อมูลการหยั่งธรณีหลุมเจาะในหน่วยฟุต ข้อมูลธรณีหลุมเจาะแบบรังสีแกมมาและข้อมูลธรณีหลุมเจาะแบบศักย์ไฟฟ้าเกิดเอง

ภาควิชา ธรณีวิทยา ลายมือชื่อนิสิต .....

สาขาวิชา ธรณีวิทยา ลายมือชื่อ อ.ที่ปรึกษาหลัก.....

ปีการศึกษา 2562 ลายมือชื่อ อ.ที่ปรึกษาร่วม.....

5632739423 : MAJOR GEOLOGY

KEYWORDS : SYNTHETIC WELL LOG / PHOTOELECTRIC LOG / MACHINE LEARNING

KAMONPHAN PHANNITHIPRASERT : RECONSTRUCTION OF SYNTHETIC WELL LOGS FROM KANSAS, USA USING MACHINE LEARNING APPROACH. ADVISOR : ASSOCIATE PROFESSOR DR. SANTI PAILOPLEE, Ph.D., 56 pp.

Photoelectric (PE) logging data is important in petroleum exploration due to its petrophysical implications, which can directly infer the reservoir composition. For example, the PE value of calcite is ~5 b/e, which can be used to indicate carbonates in the reservoir. However, well logging requires significant financial resources and intensive labor to acquire necessary information. Moreover, missing data at depth is a common problem during well logging surveys. This study thus aims to use three machine learning models: Extreme gradient boosting (XGBoost), Support vector regression (SVR), and Artificial neural network (ANN) to synthesize the PE log in the Anadarko basin, Kansas, USA. Over 50,000 well logging data points of 6 logging types (gamma ray, deep resistivity, spontaneous potential, density porosity, bulk density and photoelectric) from 12 wells are used to train, validate, and test the models in the ratio of 70:20:10. ANN performs poorly and shows the highest MSE at 0.197 due to its sensitiveness to imbalanced data. XGBoost shows the lowest mean square error (MSE) at 0.139 and R-square at 0.75, suggesting that XGBoost outperforms SVR and ANN. This is because XGBoost has an ability to handle imbalanced data, prioritize feature importance, and mimic human decision. Top three important features for synthesizing the PE log include depth, gamma ray log, and spontaneous potential log.

Department :	Geology	Student's Signature.....
Field of Study :	Geology	Advisor's Signature.....
Academic Year :	2018	Co-advisor's Signature.....

Acknowledgements

This project is accomplished with a lot of assistance. There are abundant people and environments that I would like to express thanks for.

First, I would like to express my gratitude to the kindness of the Data Resource Library, the Kansas Geological Survey for providing well logging data utilized in this study.

Secondly is my advisor and co-advisor, Associate Professor Dr. Santi Pailoplee and Assistant Proferssor Dr. Waruntorn Kanitpanyacharoen respectively. It is a whole-hearted expression that your dedicated time and advice are proved to be a major part toward the success of my project. They also support me both in academic and living problems.

Next, I would like to pay special thanks to Mr. Worapop Thongsame and Mr. Thanyaboon Sudhasirikul. They are my seniors who get involved in helping me accomplish this project. Regardless of whether the problem is difficult or easy, they both kindly assist me fix those problems.

Lastly, I would like to thank all of the Department of Geology staffs for providing support and places for me during this study.

Kamonphan Phannithiprasert

Author

List of Contents

List of Figures

List of Tables

Chapter 1 Introduction

1.1 Introduction

Subsurface information is valuable knowledge in petroleum and geothermal exploration. Geophysical well logging becomes an important tool in determining subsurface data as it records *in-situ* physical rock properties. Well logging can be used to determine lithology, fluid types, density, and porosity. Various types of logging are typical in exploration industries due to their high sensitiveness to lithology, porosity, and fluid saturation. For example, the spontaneous potential (SP) log and gamma ray (GR) log. In terms of lithological discrimination, the gamma-ray log is widely used as a shale-index indicator since its ability to detect naturally emitted gamma-ray generated from K, Th, and U elements in the formations. The gamma-ray log is thus normally used to differentiate between clay-rich and clay-poor rocks. In addition, the photoelectric (PE) log is sensitive to the mean atomic number of minerals in the formation, but not sensitive to porosity and fluid saturation presented in the rocks (Glover, 2012). The PE log is commonly used to distinguish the ambiguities of lithologies in multi-mineral formations.

A study by Dubois et al. (2007) uses well logs from the Panoma fields in Kansas to determine lithofacies comparing between conventional statistical models and machine learning methods. Two experiments were taken in both methods where the first experiment is to determine lithologies by using well log data including PE logs. Another experiment is to determine lithologies by using well log data excluding PE logs because the PE logs are not available in all wells. Results show that machine learning approaches are more accurate (up to 70%) than those of the conventional models (up to 60%). The average accuracy improves approximately 2.5% when PE logs are taken into account. The study implies that the PE log is a powerful lithology discrimination tool applicable to both human and machine-based interpretations. A study by Giao and Chung (2017) focuses on the characterization of carbonate reservoirs in Red River basin, Vietnam. The study uses the density porosity logs and the PE logs to compute apparent dry grain density (D_{ga}) and apparent volumetric factor (U_{ma}) cross plot. The D_{ga}-U_{ma} cross plot is proved to successfully identify predominant minerals in each lithological zone. The implication of the study is that the PE log is also one of the robust parameters for carbonate zonations.

Well logging requires significant financial resources and intensive labor to acquire necessary information. Moreover, missing data at depth is a common problem during well logging surveys due to the limitation of technique and complexity of the formation (Chopra et al., 2005; Rolon et al., 2009). One novel approach to remediate this problem is to reconstruct the synthetic well logs from the existing logs, using machine learning approach (Rolon et al., 2009). Machine learning is a data-driven statistical analysis technique, which is capable of learning and pattern recognition from input data in order to make predictions or discover new knowledge. Machine learning involves two types of tasks: supervised and unsupervised learning. Supervised learning is to learn or train on a function between given input and labeled outputs (Simon et al., 2015). In contrast, unsupervised learning tries to learn a relationship between data without having labeled outputs.

Machine learning has been increasingly used in statistical analysis where the other methods cannot solve high-dimension or non-linear relationships of the input data. A study by Demolli et al. (2019) performs long-term wind power forecasting based on daily wind speed data by using five machine learning algorithms: Least Absolute Shrinkage and Selection Operator (LASSO), k-Nearest Neighbor (kNN), Extreme Gradient Boosting (xGBoost), Random Forest (RF), and Support Vector Machine (SVM) algorithms. RF and SVM models are promising tools to give high accuracy prediction of wind power in different locations throughout the world with R-square values of RF and SVM are 0.995 and 0.955, respectively. In addition, machine learning is applied to energy resource exploration. For example, the bottom-hole temperature and formation temperature are predicted by using drilling fluid data in Gul et al. (2019). The study suggests that RF and xGBoost models provide high-accuracy results with R-square values more than 0.970 in both algorithms for both bottom hole circulating temperature and formation temperature. According to Goutorbe et al. (2006); Lashin (2005); Wang et al. (2013); Xie et al. (2018), well log data is utilized either in classification or regression problems. Wang et al. (2013) and Xie et al. (2018) select well logs, such as gamma-ray, deep-resistivity, and caliper logs, as input data to perform lithology classification. The results of these studies show that xGBoost provides the highest prediction accuracy but RF is more prone to overfitting. Other studies such as Goutorbe

et al. (2006) and Lashin (2005) use Artificial Neural Network (ANN) to solve regression problems. ANN is robust in both works with a correlation coefficient of more than 98 percent. A study from Puskarczyk (2019) analyzes electrofacies of Polish paleozoic shale gas formations through a set of log responses to determine reservoir potential by unsupervised learning method. Clustering analysis, SVM, and self-organized Artificial Neural Network named Kohenen neural network are used to classify the standard well logging data; gamma-ray, deep-resistivity, neutron porosity, compressional wave slowness, bulk density, and photoelectric logs. The results can distinguish nine electrofacies or clusters with only partly overlap area, suggesting that the input log data provides good information about lithology, porosity, and saturation

. In addition, photoelectric factor, high-resolution acoustic, sonic, and density logs were synthesized by various algorithms, giving high accuracy results with low mean square error (MSE) (Akinnikawe et al., 2018; Zhang et al., 2018).

Machine learning has been proven as a powerful tool to solve complex problems with high accuracy in a short amount of time. This study thus aims to reconstruct a synthetic photoelectric log from Kansas, USA by using three machine learning algorithms, particularly Artificial Neural Network (ANN), Support Vector Machine (SVM), and Extreme Gradient Boosting (xGBoost). Synthetic well logs can be created with high accuracy and scaled over a deep range of the interested area at a much lower cost and shorter time than the traditional well logging method. Kansas is located in the midwestern part of the USA and hosts a few important petroleum basins such as Hugoton and Sedgewick basins. The Hugoton area is the largest gas field in North America and one of the largest gas fields in the world, which over the last decade cash receipts form Kansas oil-and-gas production are identical to total annual statewide crop production (Kansas Geological Survey, 2001). Well log dataset is obtained from the Kansas Geological Survey website in form of Log ASCII Standard (LAS) files. The wells are located in the Hugoton Embayment, an extension of the Anadarko basin underlies in the southwestern part of Kansas state, in which the embayment axis plunges southeastward deepening into the Anadarko basin.

1.2 Objectives

1.2.1 To reconstruct synthetic well logging responses from Kansas, USA using machine learning models

1.2.2 To compare the performance of three machine learning models: Extreme Gradient Boosting (xGBoost), Support Vector Machine (SVM), and Artificial Neural Network (ANN)

1.3 The scope of study

This study aims to reconstruct a synthetic photoelectric log through well-logging data retrieved from Kansas, USA by using three machine learning algorithms, particularly Artificial Neural Network (ANN), Support Vector Regression (SVR), and Extreme Gradient Boosting (xGBoost). Twelve wells drilled in Hugoton embayment, Anadarko basin are used which are L.Maune 6, Young 'J' 1, Lighty 33-1, Nina Seyb 1, MLP SCOTT 'A' 1, J-G Unit 1-13, HCU 2220-B, Calkins 15-1, Kysar 1-1, Vercimak 'A' 1, PATTERSON UNIT 4-25, and Prentice 'A' 1. Each well is collected associated with six common well logging types: gamma ray (GR), photoelectric factor (PE), deep resistivity (RT90), spontaneous potential (SP), density porosity (DPHI), and bulk density (RHOB). A total number of the well-logging data points is 51,385 which 70% of the data points are used as a training set, 20% are used as a validation set, and the remaining are used as a test set.

1.4 Expected results

1.4.1 The photoelectric synthetic log reconstructed from well-logging data acquired from Hugoton embayment, Anadarko basin, Kansas by three machine learning algorithms

1.4.2 Model performance comparison offering information of which algorithms can provide a satisfying result in reconstructing well log measurements

Chapter 2 Literature Reviews

2.1 Well logging

Lithological information such as the composition, texture, and structure of the rocks, is necessary for determining the type of hydrocarbon reservoir. In general, hydrocarbon reservoirs can be categorized into clastic (sandstone) or carbonate (limestone and dolomite) reservoirs. There are three main techniques to collect subsurface information for reservation characterization. First, geologists can collect cores or subsurface rocks directly while drilling, The second technique is to collect cuttings flushed to the surface during drilling, and the last one is to interpret from well logging. Cores are the most valuable and accurate subsurface data yet they are not practical for economical and technical reasons. Cuttings, fragments of rocks, are the second main source of subsurface data, but they have experienced mixing, leaching, and contamination along the way they transported up to the surface. Therefore, well logging, the last technique, is an important tool to collect subsurface data *in situ* (Serra, 1984). Well logging used in this study is shown in Table 2.1.

Parameter	Nomencalture	Unit	Primary uses
GR	gamma ray log	API	Shale content estimation in the formation
RT90	deep resistivity	ohm-m	Fluid types determination in the formation
SP	spontaneous potential	mV	Porous and permeable zones determination in the formation
DPHI	density porosity	%	Formation's porosity estimation by calculating log responses from density tool
PE	photoelectric factor	b/e	Indication of major minerals in the formation
RHOB	bulk dnsity	g/cm3	Formation's porosity estimation, normally considered with NPHI

Table 2.1 Six well logging types or features that are used in this study

2.1.1 Gamma ray

Gamma ray (GR) log is a continuous measurement of gamma ray decayed from radioactive elements such as K, U, and Th in rocks. K and Th seem to be concentrated in clay minerals and U is usually rich in source rocks since its ability to be absorbed by organic matter (Puskarczyk, 2019). The obtained GR responses are the summarized contribution of all three elements. The API (American Petroleum Institute) is applied as a measurement unit in most cases according to Tiab et al. (2012) stated that one API equals 0.07 micrograms of radium equivalent per ton of formation. In general, An important application of the GR log is to calculate the shale content with the assumption that only shale contains radioactive minerals in a rock, for instance, K-feldspar (Schön, 2015). As shown in Figure 2.1 that high gamma ray is shale and low gamma ray is sandstone in the example, but it can be other rocks such as limestone depending on depositional environments.

2.1.2 Deep resistivity

Resistivity, an inversion of conductivity, measures electrical resistances of the formation. It is the primary property to determine reservoir properties especially porosity and fluid types (Thongsame, 2018). The coverage range of formation resistivity is widely distributed from 0.1 to 1000 ohm-m where the rocks with resistivity higher than 1000 ohm-m are interpreted to be very impervious or very low porosity (Schön, 2015). Noteworthy that the resistivity resulting log presents in three curves: shallow, medium, and deep corresponding to the radius of investigation (Figure 2.1). Deep resistivity is presumed to true formation resistivity and separation between shallow and deep curves can be used to determine the fluid type, the diameter of mud invasion, and zones of permeable rocks (Varhuag, 2016). Figure 2.1 suggests that if the separation between shallow and deep curves are wide, it tends to contain hydrocarbon. In contrast, if they have narrow separation or move in the same trend, it is a water-bearing zone.

2.1.3 Spontaneous potential

Spontaneous potential (SP) log is a measurement of electrical potential occurring naturally, which is produced by the exchange of fluids with different salinities between the

formation water and the borehole mud filtrate. During drilling permeable bed is invaded by mud fluid causing ions exchange. If the formation fluid has more salinity than the filtrated mud, then the SP curves response to the left arbitrary to the SP baseline generated from impermeable shale formations. On the other hand, the SP curves will deflect to the right if the formation fluid has less salinity than the filtrated mud. In the latter case is usually a water-filled permeable formation. So, the SP curves can indicate which formations are permeable (Varhuag, 2016).

2.1.4 Bulk density

Bulk density (RHOB) log is a key measurement of porosity since porosity, in general, is inversely related to the rock density. This log is obtained by density tool which emits gamma rays to the formation. Compton scattering occurs when the emitted gamma ray collides with electrons inside the formation and gives off energy. Thus, the number of collisions directly implies the number of electrons in the formation. The tool counts scattered gamma rays reaching back to the detector. Low-density formations, high porosity, will get more counts on the number of coming back gamma rays than high-density formations (Varhuag, 2016). Apart from that RHOB log can detect organic-rich shales since the organic matter has a much lower density than non-organic-bearing shales (Puskarczyk, 2019). As illustrated in Figure 2.1, shales have a higher density than sandstones.

2.1.5 Density porosity

Density porosity (DPHI) log is another form of porosity calculated from density data acquired from the density log using the equation below. Where: ϕ is calculated porosity, ρma is matrix density, ρb is formation bulk density (reads from log values), and ρf is fluid saturation density.

$$\Phi = \frac{\rho_{\mathrm{ma}} - \rho_{\mathrm{b}}}{\rho_{\mathrm{ma}} - \rho_{\mathrm{f}}}$$

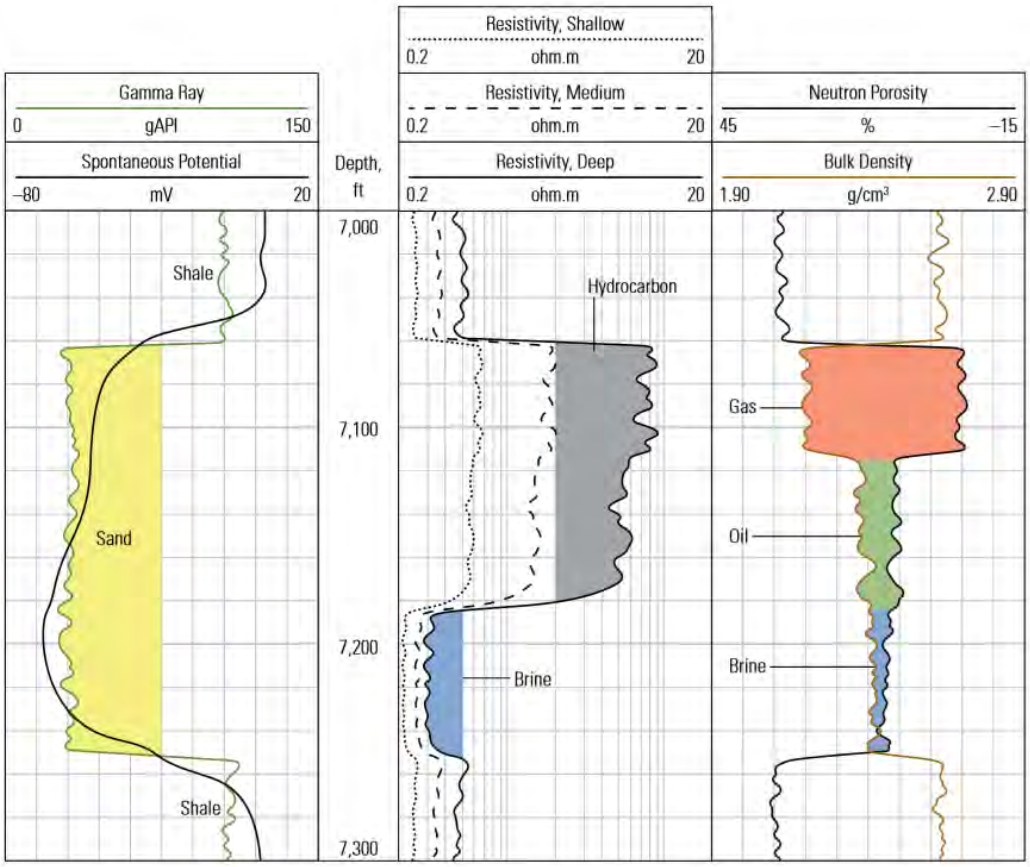


Figure 2.1 A basic suite of logging measurements. Track 1 is gamma ray log and spontaneous potential log often used as lithological classification. The next column is called a depth track (in this diagram represented in feet). Track 2 is resistivity measurements used in determining the fluid types. Track 3 is a neutron porosity log and bulk density log used to estimate porosity (Varhuag, 2016).

2.1.6 Photoelectric factor

Photoelectric (PE) effect is a phenomenon where the gamma photon energy emitted from the density tool interacts with an electron in the rock. Innermost-shell electrons are ejected from the atom and the gamma photons are absorbed. Higher shell electrons will release energy to fill the vacancies in the innermost shell in the form of X-ray or low-energy gamma ray. Unlike the Compton scattering process that the photon energy transfers only part of its energy to the electrons and scatters away. Photoelectric factor (PE) log is also earned from the density tool recording low-energy gamma rays reaching the detector. The aggregate atomic number (Z) of the

elements assembled in the formation is a direct function of its recording, thus it is a precious log for mineral determination in the formation (Asoodeh and Shadizadeh, 2015). In addition, porosity and fluid saturation in the rocks are considered negligible (Figure 2.2). One of the exceptions is highly saturated brines which give a significant PE value (Glover, 2012).

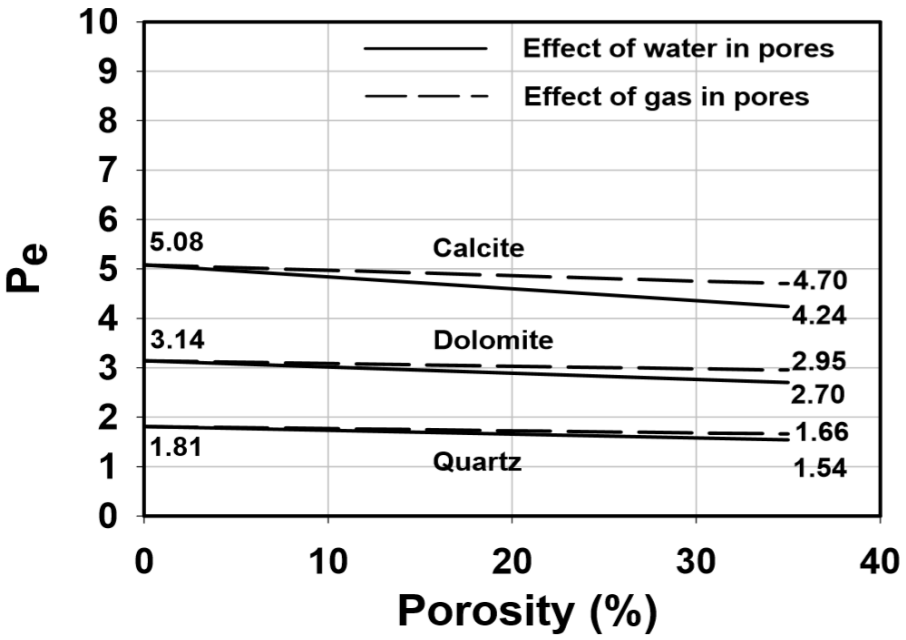


Figure 2.2 PE value as a function of porosity and fluid saturation. Notice that PE values change insignificantly, even the porosity is increasing in all mineral types. The fluid saturation influencing PE values are disregarded as well (Glover, 2012).

2.2 Machine learning algorithms

Machine learning (ML) is a technology grown out of artificial intelligence (AI) work. The goal of machine learning is to enable machines to perform excellently by using intelligent software whose backbone is developed from statistical learning methods. Mitchell et al. (2006) stated in his article entitled "The Discipline of Machine Learning" that this discipline seeks to answer the question of "How can we build computer systems that automatically improve with experience, and what are the fundamental laws that govern all learning processes?". He also precisely defined that machine learning is a computer program that learns from experience **E** with respect to some task **T** and some performance measured by **P**. There are two main machine learning techniques:

supervised learning and unsupervised learning. In supervised learning, it is a label-based technique that requires both input independent variables and labels as the training data. Two categories under the supervised learning field discriminate by the data type of labels. The first group is called regression problem for numerical labels, another is called classification problem for discrete or categorical labels. In contrast, unsupervised learning requires only the input data with no label, thus it is appropriate for cluster discovery which clusters from the relationship between data in the input data. After the machine learning model has learned from the training data, performance assessment is needed. The unseen data, so-called the test data will be given to the model and evaluated by different evaluation metrics. In regression problems, evaluation metrics can be R-square, root mean square error (RMSE) or mean square error (MSE). Classification problem, on the other hand, evaluation metrics can be confusion matrix, area under curve, or F1-score. Note that in this study three supervised learning models namely; extreme gradient boosting (XGBoost), support vector regression (SVR), and artificial neural network (ANN) will be used to solve regression problems and assessed by R-square and MSE evaluation methods.

2.2.1 Extreme Gradient Boosting (XGBoost)

Extreme gradient boosting (xGBoost) is a scalable implementation of the gradient boosting technique. Gradient boosting is a predictor that ensembles many weak learners together, typically decision trees (Biau and Carde, 2017). A decision tree is non-parametric supervised learning used for both classification and regression tasks, in which the algorithm works as a rule-based system (if-then-else rules). The decision tree architecture is shown in Figure 2.3. The feature that best separates the training data into two subsets will be used as a root node and do the same for the child nodes. After done with all the splitting, the last node (with no child node) is then the prediction node, called leaf nodes.

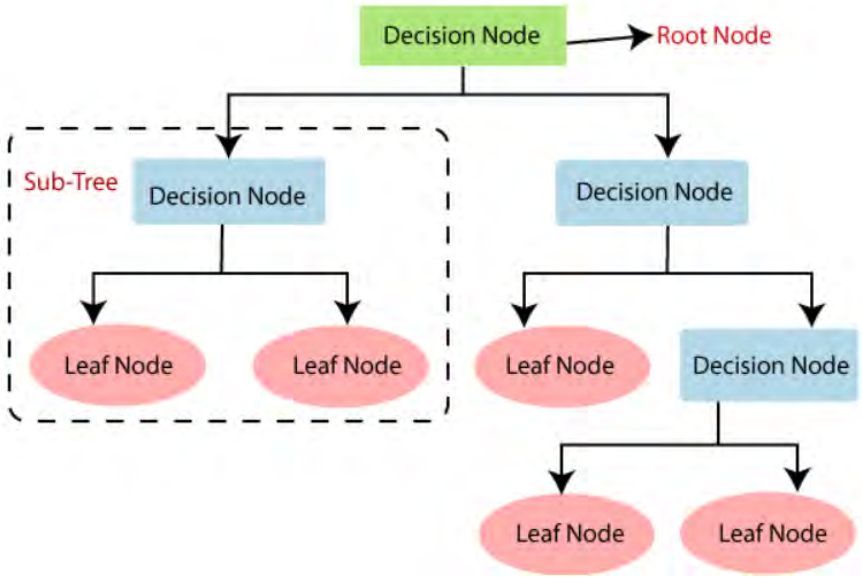


Figure 2.3 The architecture of the rule-based decision tree model, where a decision node in the green box is the root node and subdivided into two child nodes (decision nodes in the blue boxes). The last nodes without further separation are the leaf nodes where the predictions are made (JavaTpoint, 2018).

However, a single decision tree suffers from the data having several features and classes and occasionally encounters overfitting issues (Aha et al., 1998; Thongsame, 2018). This where boosting method becomes handy. xGBoost trains many trees in gradual, additive, and sequential manners since the remnant errors in the previous tree are used as labels in the next tree (Figure 2.4). Thus, in every tree added to the algorithm, the algorithm focuses on the error labels and attempts to correct them resulting in a better model in each step. For xGBoost, *learning rate* and *maximum depth* parameters were optimized. *Learning rate* controls how fast a tree, added to xGBoost sequentially, learns to correct the residual errors from the previous tree. *Maximum depth* controls the maximum depth of each tree fed to the xGBoost and increasing of this value can lead to overfitting (Zhang and Zhan, 2017).

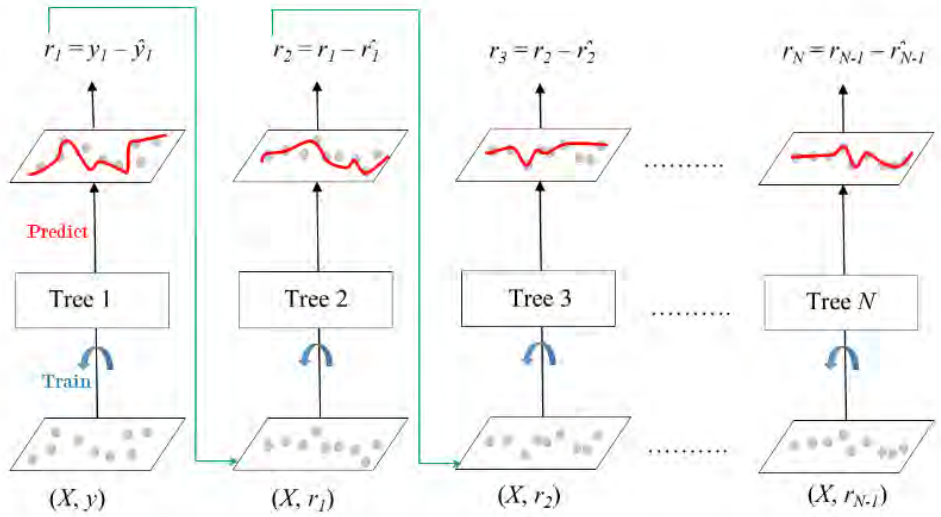
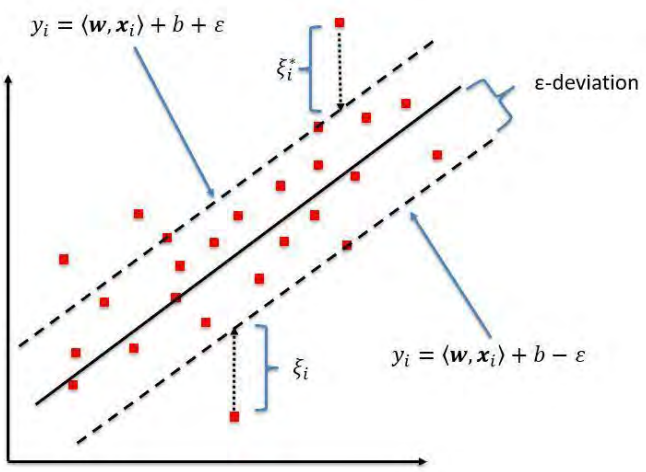


Figure 2.4 A schematic diagram of how the boosted tree works, showing input data (X), true labelled output data (y), prediction (y-hat), the previou tree remnant error (r), and the current tree remnant error from the prediction (r-hat) (Kawerk, 2018).

2.2.2 Support Vector Regression (SVR)

Support vector regression (SVR) is a kind of support vector machine (SVM) used for regression analysis. Aiming of the algorithm is to find the best function to estimate the output. The theory is developed on a basic idea of a regression algorithm as well as a linear product of two vectors in Hilbert space, space where the linear product has a real value (Maleki et al., 2014). The difference between a simple regression and SVR is that the regression algorithm tries to minimize the training error using all of the training examples, yet SVR tries to minimize the generalized error bound using a subset of the training examples. The generalized error bound is the combination of the training error and a regularization term controlling model's complexity (Basak et al., 2007). *C* is a regularization term indication of how much the model can tolerate misclassifying of each training example outside the boundary decision (Figure 2.5). Higher *C* means the model cannot tolerate any misclassified data leading to overfitting. In contrast, lower *C* can tolerate many misclassified data leading to underfitting. To optimize generalization of the SVR model, errors within a certain distance, *epsilon* or a margin of tolerance, are ignored and SVR uses

only values outside an epsilon-intensive tube to construct the model (Kleynhans et al., 2017). Thus, *C* and *epsilon* were tuned in this study.



$$\text{minimize} \quad \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{n}(\xi_i^* + \xi_i)$$

$$\text{subject to} \begin{cases} y_i - \langle \mathbf{w}, \mathbf{x}_i \rangle - b \leq \epsilon + \xi_i^* \\ \langle \mathbf{w}, \mathbf{x}_i \rangle + b - y_i \leq \epsilon + \xi_i \end{cases}$$

Figure 2.5 One-dimensional example of SVR where the solid line is a hyperplane used to predict the target value (y). The two dashed lines are boundary lines that are epsilon distances away from the hyperplane (Kleynhans et al., 2017).

Figure 2.5 is a simple linear SVR model, but in case of non-linear SVR model, the data points will be mapped into a high dimensional space called feature space by a kernel function (Kleynhans et al., 2017). Radial basis function kernel (RBF) is used in this study and also in Kleynhans et al. (2017), Maleki et al. (2014), Thongsame et al. (2018), and Xie et al. (2018). Radial kernel behaves as a weighted nearest neighbor model which means the closest data point or nearest neighbor influencing the most on prediction of a new observation.

2.2.3 Artificial Neural Network (ANN)

In 1980, an artificial neural network (ANN) was first introduced by Kunihiko Fukushima which is inspired by modeling the human brain (Simon et al., 2015). ANN is an interconnected network that contains a collection of neurons that mimic problem-solving skills of the brain by learning, developing, and establishing a mathematical approximation for non-linear patterns between input and output data (Buhulaigah et al., 2017; Chitsazan et al., 2015; Prieto et al., 2016). ANN can be categorized by the direction of the input data into feedforward and feedback neural networks.

Feedforward neural networks are defined when the flow of the signals is fed only in one direction. If the neural network has some kind of internal recurrence, the flow of signals can be fed back to the previous neuron or layer, then it is a feedback neural network. The multilayer perceptron feedforward neural networks are used often in generating synthetic well logs (Akhundi etal., 2014; Melaki et al., 2014; Onalo et al., 2018; Ramcharitar and Hosein, 2016; Rolon et a., 2009; Zoveidavianpooretal, 2013). A multilayer perceptron is a fully connected multilayer feedforward supervised learning consisting of three layers: an input layer, hidden layers, (can be more than one) and an output layer. Figure 2.6 illustrates an example of a multilayer perceptron that all connected lines point in only one direction, no loops exist in the network.

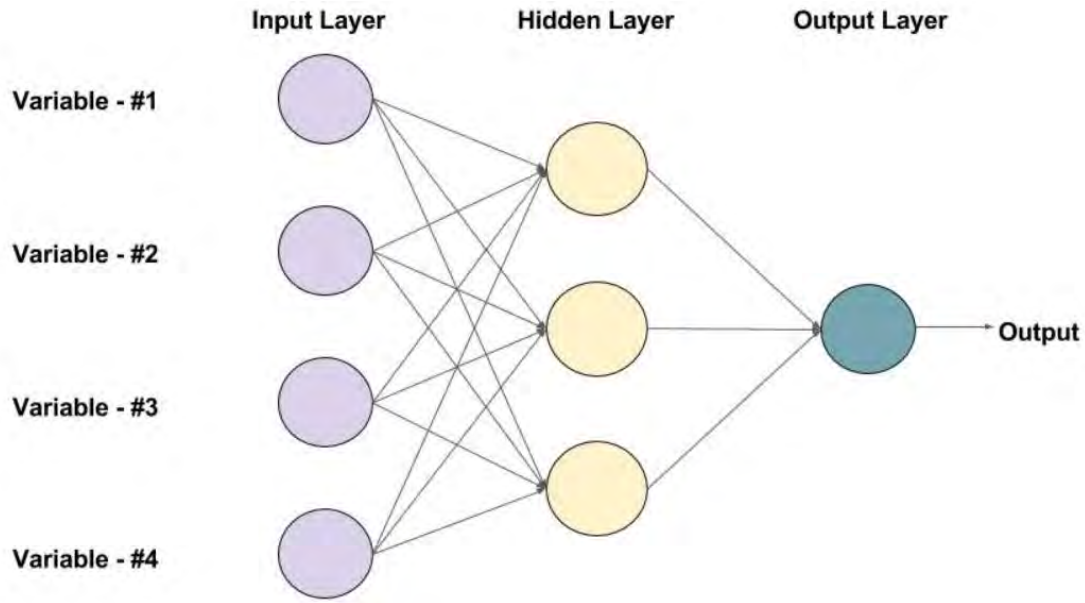


Figure 2.6 An example of a multilayer perceptron with one hidden layer (with three Qneurons) (Gupta, 2019).

Each node in the layer is a neuron which is the basic unit of a neural network. The first layer is an input layer that provides input data or features to the network where each node represents each feature. In the hidden layer is where the calculation begins inside each hidden node, as shown in Figure 2.7, there are two steps involved.

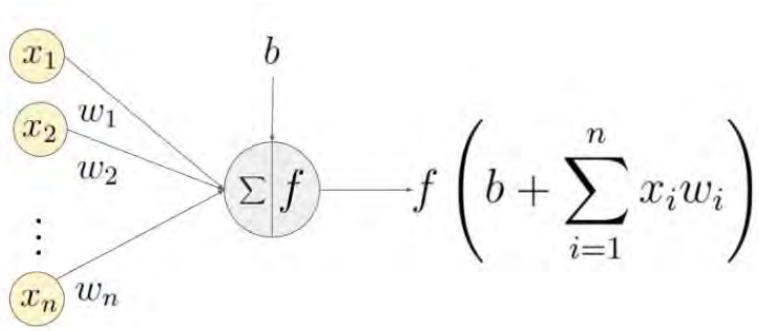


Figure 2.7 An example of a node, neuron, with input data or features (x_1 - x_n), their associated weights (w_1 - w_n), a bias (b) and the activation function (f) are applied to the weighted sum of the input (Gupta, 2019).

First, it calculates the weighted sum of its input corresponding to the weights associated with each input (x_1w_1, x_2w_2 ,..., x_nw_n), noteworthy that these weights are parameters the algorithm tries to learn in the training phase. Then the output from the first step is applied to the activation function, which can be linear or non-linear to normalize the output. For the last layer, output layer, it is where the prediction given which the activation function here can differ depending on the problem whether classification or regression. The *number of nodes* in the hidden layer and the *dropout* are adjusted in ANN. *Rate of dropout* is used to prevent overfitting by setting a chance to some number of nodes to zero or drop that node out while learning.

Chapter 3 Study area

Anadarko basin is the deepest Phanerozoic cratonic basin in the USA covering an area of 70,000 square miles (130,000 km^2) and 40,000 ft (12 km) in thickness, which also one of the greatest oil-and-gas producing provinces in the Northern America craton (Kansas Geological Survey, 2001). In this area, Oklahoma is the first formed basin, which will form the Anadarko basin in the latter time. Pennsylvanian is the time when the inversion normal faults occurred and resulted in the Wichita uplift and the Anadarko basin (Tomlinson and McBee, 1959; Ham et al., 1964; Ham and Wilson, 1967). The basin center is situated in western Oklahoma and northern Texas, extending to southwestern Kansas and southeastern Colorado. The basin is bounded by the Las Animas arch to the west, the Namaha uplift to the east, the Central Kansas uplift to the north, and the Wichita-Amarillo uplift to the south. The basin evolution history can be divided into four stages (Johnson, 1988), viz. 1) An igneous stage, 2) An early epeirogenic stage, 3) An orogenic stage, and 4) A late epeirogenic stage.

3.1 An igneous stage, Precambrian to middle Cambrian periods

This stage is represented by Precambrian and Cambrian igneous and metasedimentary basement rocks, which are considered to underlie the Anadarko basin. Evidenced by seismic and drilling data gathered from the eastern, western, and northern flanks of the basin instead of the unreachable deeper southern part. From the drilled data, the Precambrian basement rocks are identified to be massive allotriomorphic granular texture (mesozonal) granitic and related rocks with ages around 1,300 – 1,600 Ma (Denison et al., 1984). During early to middle Cambrian, rifting had occurred associated with the proto-Atlantic Ocean opening (Burke and Dewey, 1973). Resulting in aulacogen, failed arms of triple junctions, development extended into the North America craton (Figure 3.1a). the failed-arm rifting allowed massive amounts of igneous to intrude shown in Figure 3.1b. Cambrian basement rocks are granites, rhyolites, gabbros, anorthosites, and basalts. The igneous Cambrian rocks are exposed in the present Wichita uplift estimated total thickness of approximately 20,000 ft (6.1 km) (Oklahoma Geological Survey, 2008).

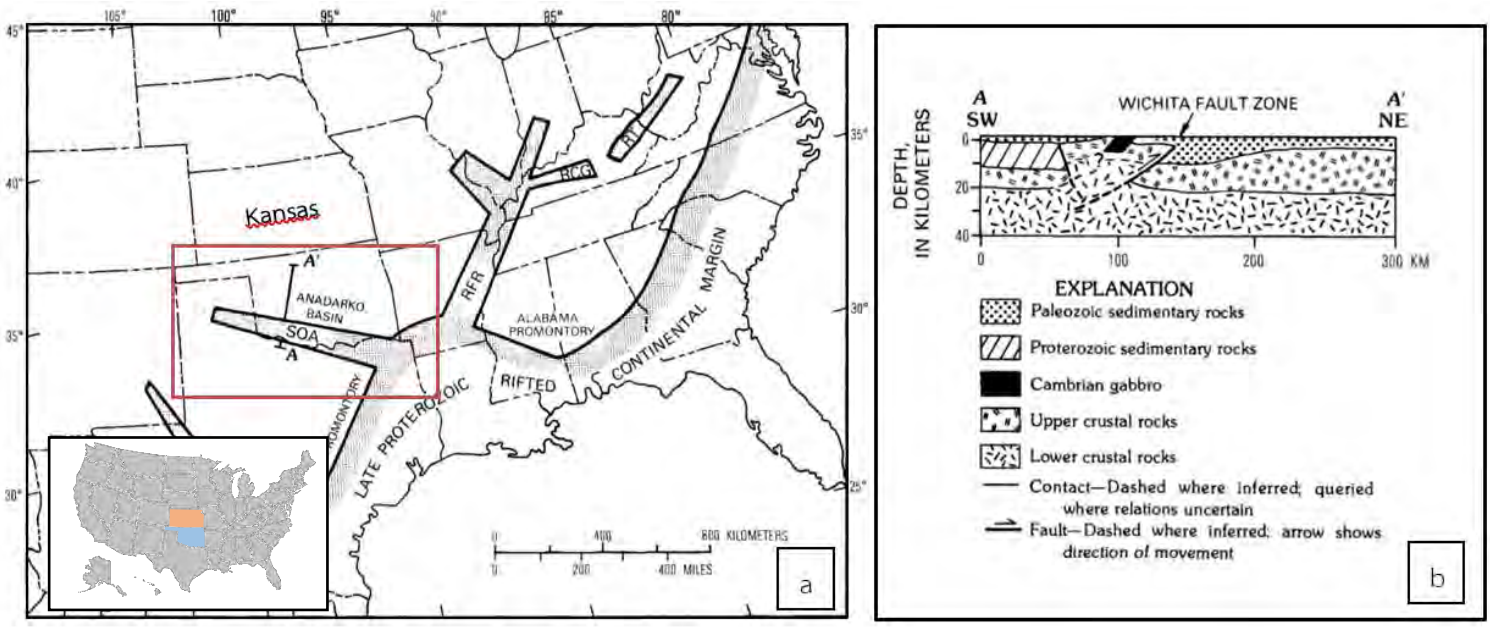


Figure 3.1 a) Oklahoma and Kansas are showing in blue and red color respectively with generalized paleotectonic map showing the southern Oklahoma aulacogen, failed arm of the triple junctions. DA, Delaware aulacogen; RCG, Rough Creek graben; RFR, Reelfoot rift; RT, Rome trough; SOA, southern Oklahoma aulacogen. b) cross section showing the geology of southern Oklahoma aulacogen, adjacent to the Anadarko basin to the north (modified from William and Perry, 1989).

3.2 An early epeirogenic stage, Late Cambrian to Mississippian periods

Early epeirogenic stage started when the aulacogen began to cool and subsidence, after the rifting phase, becoming the Oklahoma basin. The sea first invaded during this time, Late Cambrian, and moved throughout the state from the east (Oklahoma Geological Survey, 2008). Therefore, the environment in the basin started with the transgressive sandstones, Reagan Sandstone (Figure 3.2) deposited upon the basement rocks. Then, shallow-water flooded into the basin resulting in limestones and dolomites deposition, the Arbuckle Group (Oklahoma Geological Survey, 2008). The Arbuckle limestones deposited continuously until Early Ordovician making up succession more than 6,000 ft (1.8 km) along the basin depocenter (Johnson, 1988). However, the succession thinner and truncated in the northwestern part of the Oklahoma basin and the

Hugoton Embayment (Johnson, 1988). Late Ordovician rocks are characterized by the Simpson Group sandstones, the Viola Group limestones, and the Sylvan shales.

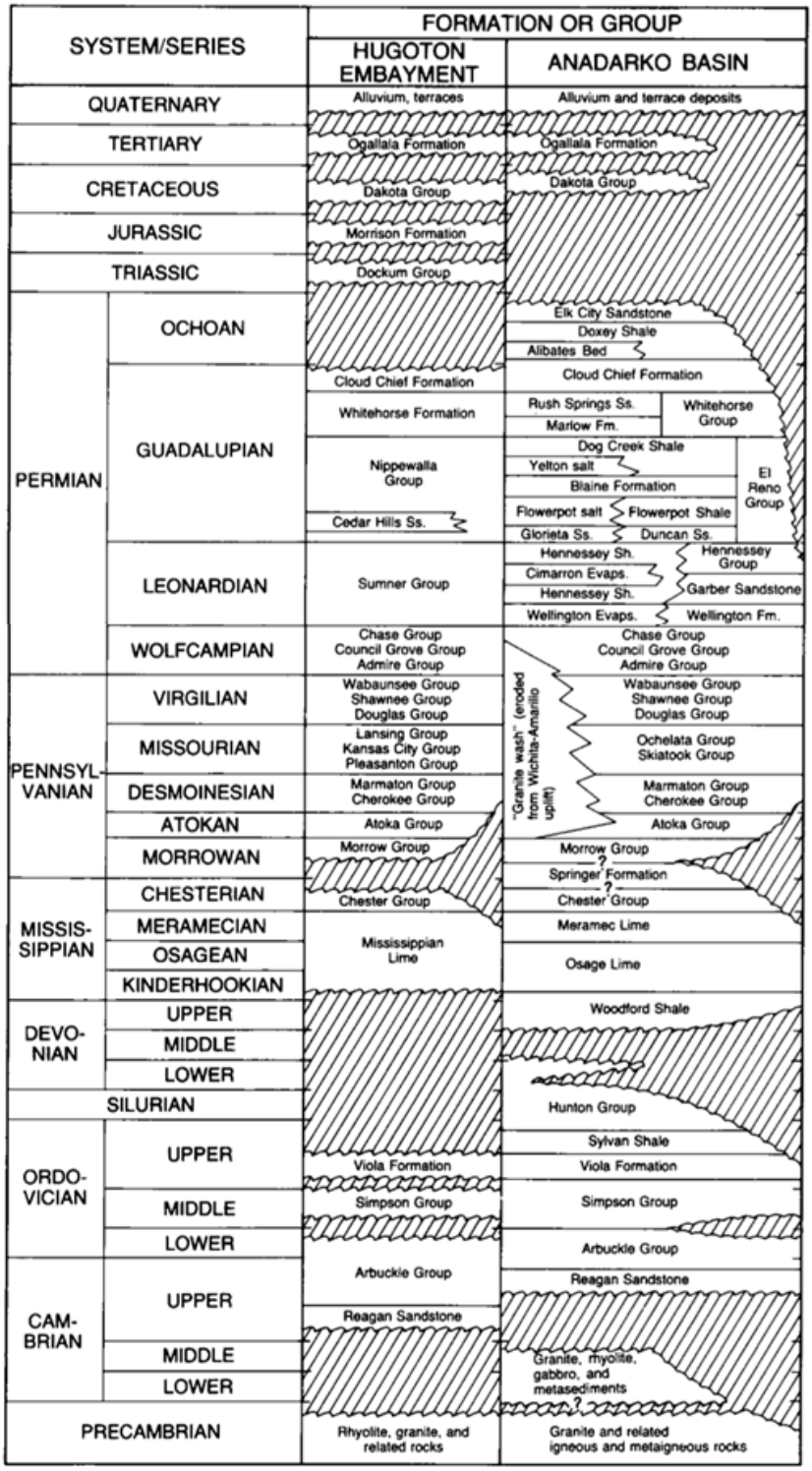


Figure 3.2 Stratigraphic column for the Anadarko basin, and the Hugoton embayment. The diagram heights are not relative to the unit thicknesses (Johnson, 1988).

In Silurian and Devonian, the Hunton Group carbonates were deposited, and overlain by the Woodford organic-rich shale (Oklahoma Geological Survey, 2008). The present Wichita uplifts contain early marine invertebrate fossils found in this period such as trilobites, bryozoans, and brachiopods (Oklahoma Geological Survey, 2008). The deposition was interrupted by two epeirogenic uplifts (Amsden, 1975): pre-middle Early Devonian, and pre-Late Devonian resulting in widespread Woodford unconformities. According to Oklahoma Geological Survey (2008), during Mississippian time, the first half of the period was dominated by limestones and cherts. Important units such as Sycamore Limestone in southern Oklahoma, and 'Mississippian lime' in northern Oklahoma, Hugoton Embayment. In the last half of the period, basin rapidly subsided in the southern part of the basin making shales dominate the area. The schematic principle rock types of each period since late Cambrian to Mississippian are presented in Figure 3.3.

3.3 An orogenic stage

This is the stage where the orogeny and subsidence occurred in the southern part of the Oklahoma basin, during Pennsylvanian, major history changes (Johnson, 1988). Prior to Pennsylvanian, proto-Anadarko basin and the Wichita-Amarillo blocks are subsided together as one (named southern Oklahoma aulacogen) (Johnson, 1988). Vast areas of the region experienced epeirogenic uplifts during Late Mississippian and Early Pennsylvanian. In southern Oklahoma aulacogen has experienced normal faults inversion (Tomlinson and McBee, 1959; Ham et al., 1964; Ham and Wilson, 1967); separating into the Wichita-Amarillo uplifts and deep Anadarko basin area. Since then, the Anadarko basin is defined and the other parts, used to be called Oklahoma basin, are restated to be part of the Anadarko basin. The Mississippian shallow-marine deposition has ended in all areas unlike the deeper part of the basin (Johnson, 1988); limestone continuously deposited (the Springer Group). The Wichita-Amarillo uplifted in N60W trends (Gilbert, 1982, 1987) with northward thrusting and associated with rapidly the Anadarko basin sinking.

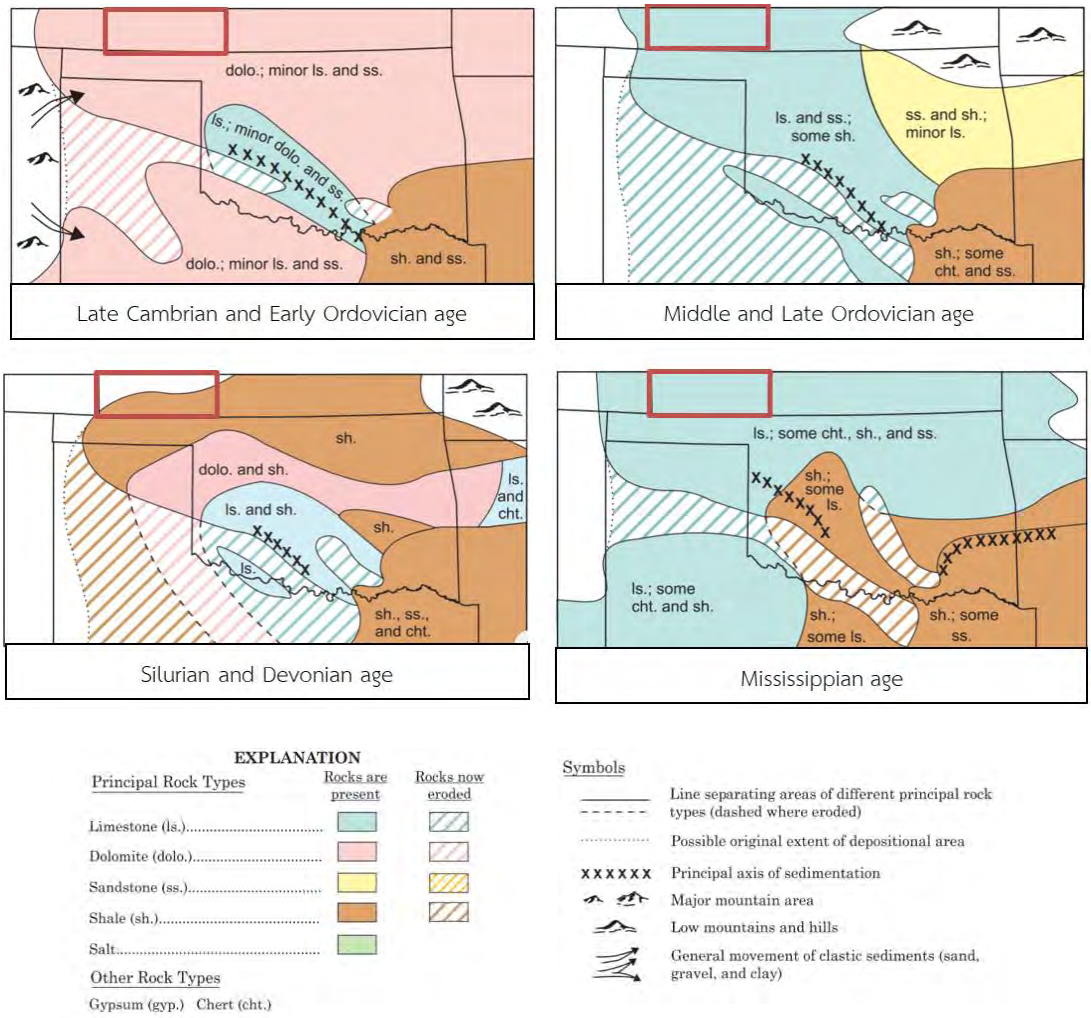


Figure 3.3 Schematic principal rock types of each period since late Cambrian to Mississippian in Oklahoma and adjacent areas. Red box shows the study area, Hugoton embayment located in southwestern Kansas (modified from Oklahoma Geological Survey, 2008).

Pennsylvanian has five epochs namely: Morrowan, Atokan, Desmoinesian Missourian, and Virgilian (all epochs evolution shown in Figure 3.4), which according to Johnson (2008) orogenies occurred in all epochs. In Morrowan and Atokan time, igneous fragments and conglomerates eroded from pre-Pennsylvanian rocks are deposited near the Wichita-Amarillo uplifts. Sandstones and shales are graded toward the basin centers. A broad north-trending uplift rose from the central Oklahoma city extended to Kansas city, and also in northeastern Oklahoma city (Oklahoma Geological Survey, 2008). During Desmoinesian (Oklahoma Geological Survey, 2008); the sea

covered all the area extending to central Kansas and some uplifts had stopped. Cyclic marine limestones and shales are dominated in the Desminesian strata with some lenticular point bars and channel filled sandstones. 'Granite wash' from the Wichita uplifts concurrently deposited on the southern part of the Anadarko basin. In Missourian and Virgilian time, Late Pennsylvanian Arbuckle orogeny. The strong compression affected many mountains in the area (Oklahoma Geological Survey, 2008). The Missourian and Virgilian strata in the Hugoton embayment, main shelf of the Anadarko basin, had accumulated marine limestones interbedded with shales. In central Anadarko basin had accumulated marine shales, and sandstones graded westward into delta facies (delta plain sandstone, and pro-delta shales) interfingering with some thick 'granite wash' along the Wichita-Amarillo uplift (Oklahoma Geological Survey, 2008).

3.4 A late epeirogenic stage, Permian till presents

By Permian period, the Wichita-Amarillo uplift has ended and subsidence with a very low rate compared with the Anadarko basin (Johnson, 1988). The Permian shallow inland sea covered most of the western Oklahoma depositing limestones and gray shales along the ancient seaway's center. Lithologies laterally grade to east and west as Permian red beds (red shales and red sandstones) (Oklahoma Geological Survey, 2008). Early Permian (Leonardian epoch), the evaporating seawater deposits salt, gypsum, and anhydrite (Jordan and Vosburg, 1963). During Late Permian, the Wichita mountain is buried by sediments from the east (Figure 3.5). Post-Permian, Mesozoic era and Cenozoic era deposit mainly red sandstones and shales mixing with fluvial, deltaic, and lacustrine deposits with lesser marine sediments (Johnson, 1988)

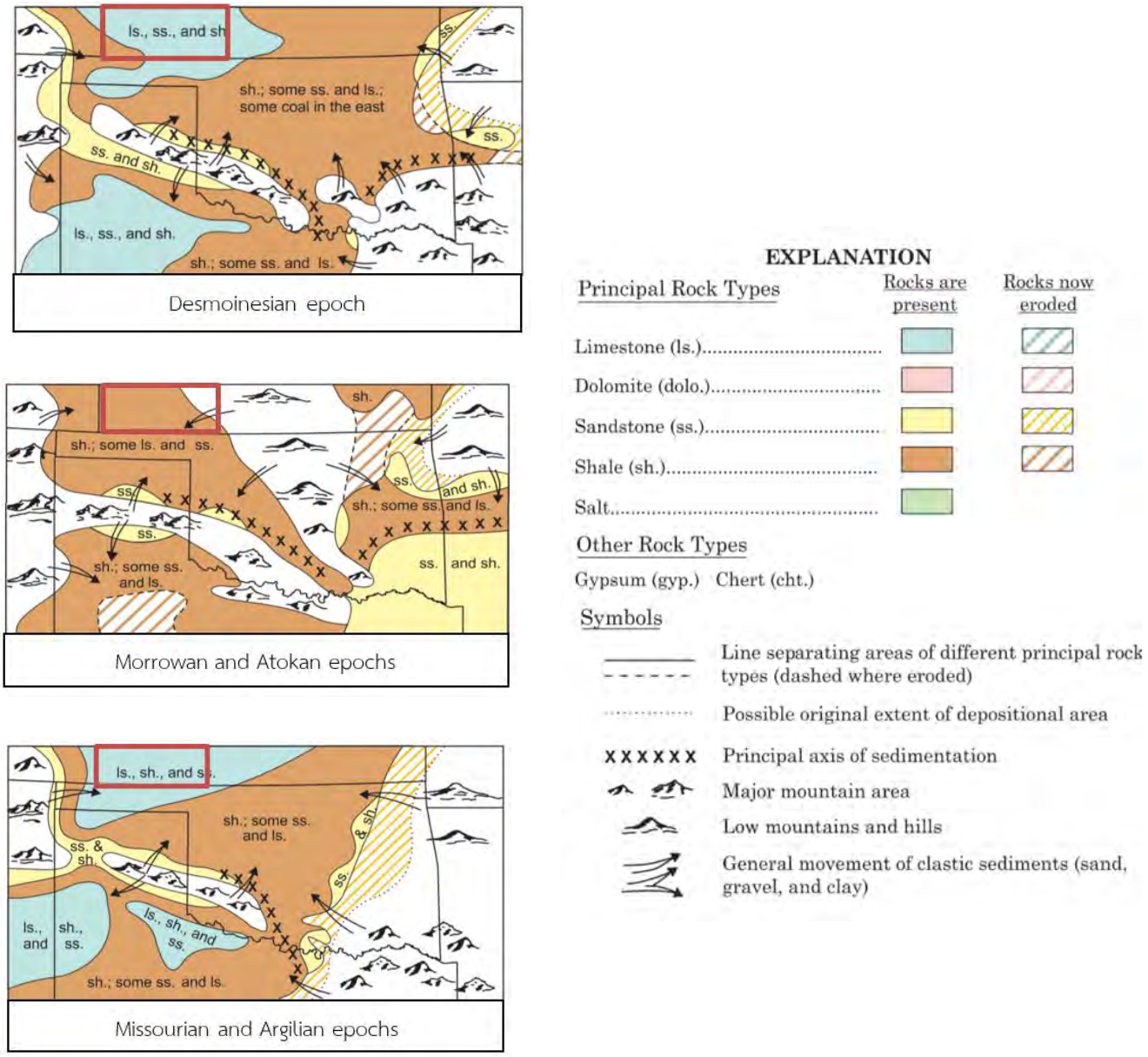


Figure 3.4 Schematic principal rock types of each epochs in Pennsylvanian period in Oklahoma and adjacent areas. Red box shows the study area, Hugoton embayment located in southwestern Kansas (modified from Oklahoma Geological Survey, 2008).

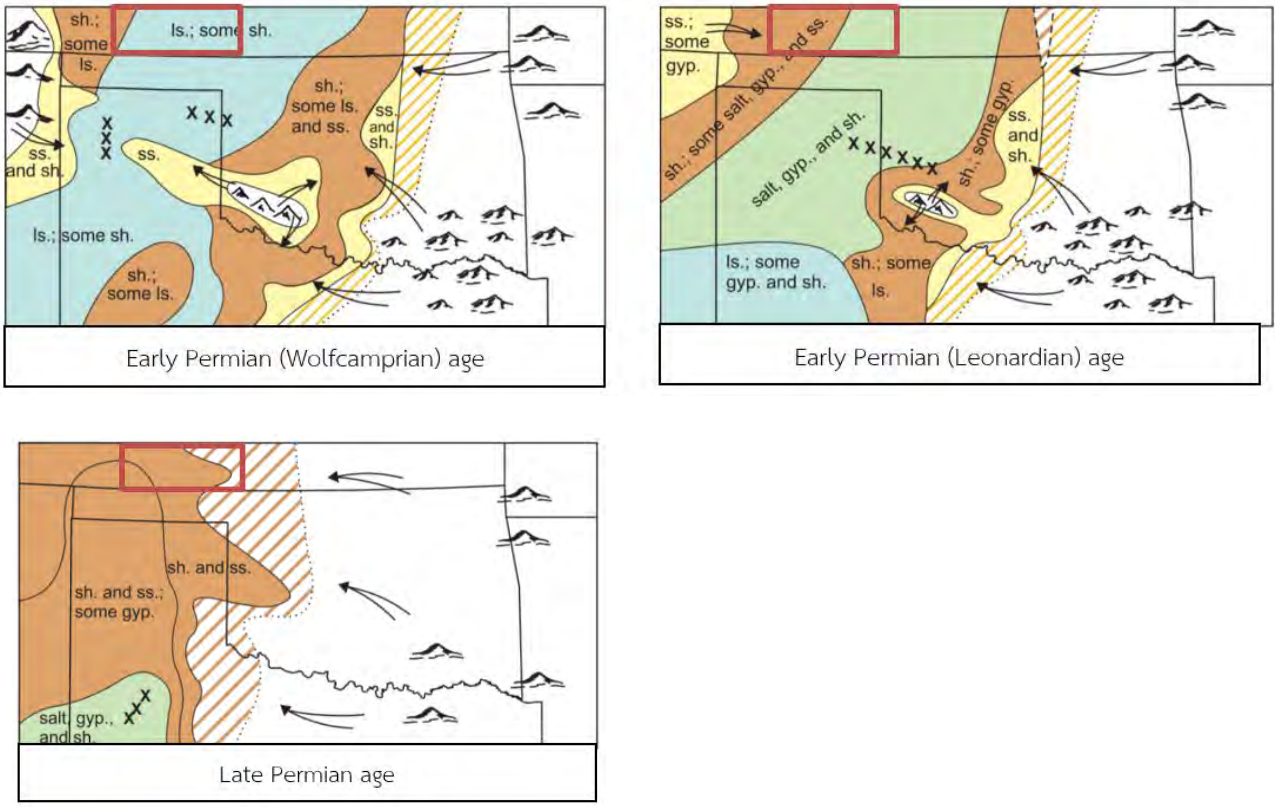


Figure 3.5 schematic principal rock types of Permian period in Oklahoma and adjacent areas. Red box shows the study area, Hugoton embayment located in southwestern Kansas (modified from Oklahoma Geological Survey, 2008). See Figure 3.4 for symbol explanations

Chapter 4 Methodology

Overview of the workflow

The procedures to generate synthetic well log are divided into four main steps: data collection, data preprocessing, synthetic well log reconstruction, and model comparison respectively (Figure 4.1).

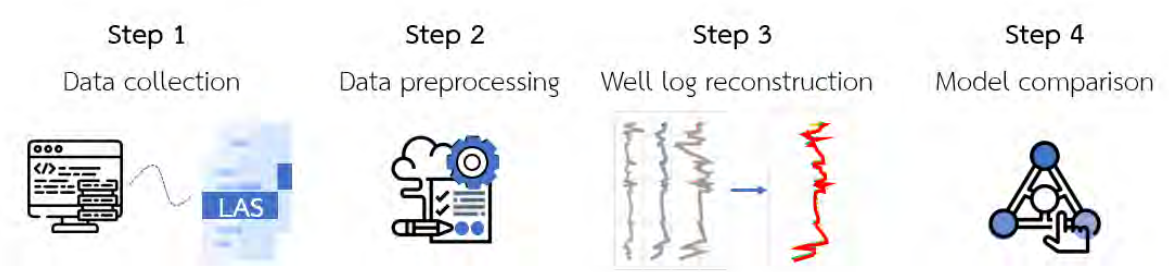


Figure 4.1 Four steps to reconstruct synthetic well log

4.1 Data collection

Well log data in this study was retrieved as Log ASCII Standard (LAS) files from Kansas Hugoton Project located on Hugoton Embayment, Anadarko basin, southwestern Kansas, USA. Kansas Hugoton Project is a five-year project funded by industry, University, and Government to develop technology and information of the Hugoton Embayment. The well data is assembled at the Data Resource Library, the Kansas Geological Survey. The data is collected to be the repository for oil, gas, and water well records in the state of Kansas available for public uses. Thus, 12 wells located on the Hugoton Embayment are selected (Figure 4.2) and 7 well logging types or features from each well are gathered: gamma ray (GR), photoelectric factor (PE), deep resistivity (RT90), spontaneous potential (SP), density porosity (DPHI), and bulk density (RHOB). The depth in each well is different, but overall ranges from 300 ft. - 5800 ft. (90 - 1800 m.). The total number of data points gained from 12 wells is 51,385 in which the data points and depth range acquired from each well are illustrated in Table 4.1. Figure 4.3 is used to demonstrate the drilled depth comparison between each well. Measurement resolution for logging tools is 0.5 ft, identical for all logging types.

Lithology data, which is contained in well report, is not provided with the LAS files, only 2 wells are possible namely, PATTERSON UNIT 4-25 (available from 3600 - 4900 ft.), and J-G Unit 1-13 (available from 4650 - 4850 ft.). These 2 well reports suggest that the lithofacies at the Hugoton Embayment at those depths are dominated by limestone interbedded with shale, and some quartz-rich sandstone also found.

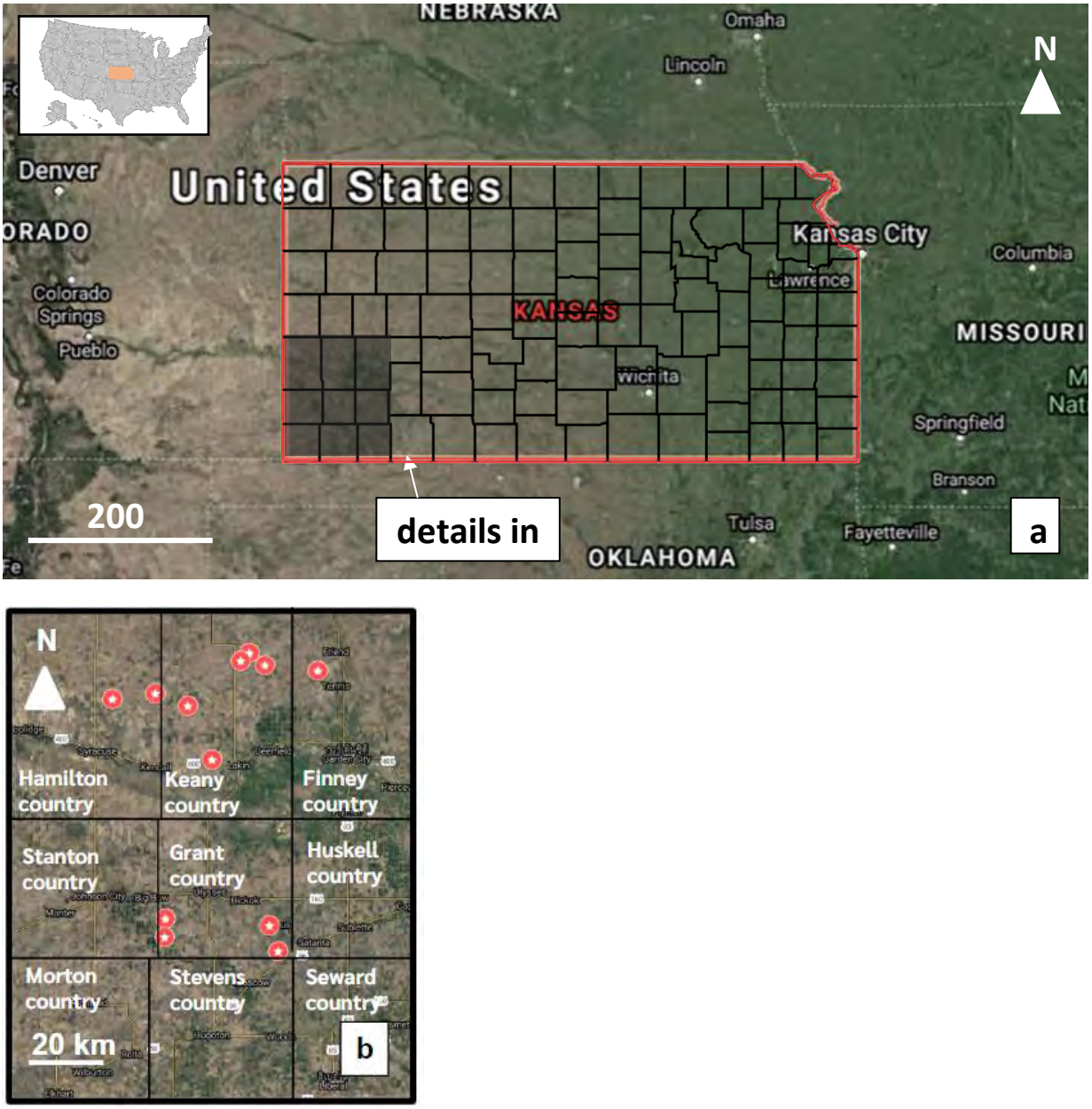


Figure 4.2 a) Kansas state locates on the midwestern part of the USA, and Hugoton embayment is laid on the southwestern part of Kansas highlighted in black. b) Locations of each well are shown in red symbols.

4.2 Data Preprocessing

This step can be divided into four steps which are data preparation, data transformation, data exploratory analysis, and data partitioning.

4.2.1 Data Preparation

Well logging data are obtained in Log ASCII Standard (LAS) file format. The data are collected in different quality and logging types. Therefore, when gathering the well data, common logging types are selected viz. Gamma ray (GR), Deep resistivity (RT90), Photoelectric factor (PE), Spontaneous potential (SP), Density porosity (DPHI), and Bulk density (RHOB).

Well name	Number of data points	Minimum depth (ft)	Maximum depth (ft)	Location (latitude, longitude)
L.Maune 6	2171	3700	4785	(38.21, -100.98)
Young 'J' 1	5769	2948	5832	(37.45, -101.52)
Lighty 33-1	5333	3048	5714	(37.49, -101.15)
Nina Seyb 1	3671	3972	5807	(37.51, -101.52)
MLP SCOTT 'A' 1	1113	5196	5752	(37.41, -101.12)
J-G Unit 1-13	8652	1051	5376.5	(38.14, -101.55)
HCU 2220-B	3977	951	2939	(38.13 , -101.70)
Calkins 15-1	6868	320	5084.5	(38.23, -101.25)
Kysar 1-1	5603	2250	5051	(38.26, -101.22)
Vercimak 'A' 1	3543	3400	5171	(37.96, -101.35)
PATTERSON UNIT 4-25	2569	3648	4932	(38.11, -101.44)
Prentice 'A' 1	2116	3845	4902.5	(38.212, -101.17)

Table 4.1 Data points and depth in ft. shown in minimum and maximum ranges from each well. Latitude and longitude where the wells were drilled also given.

The data originally consists of approximately 51,385 data points. However, the missing values are still existing in the dataset which may result from the limitation of technique and error in well log reading. Since the objective of this study is to generate photoelectric (PE) log, thus the PE log must have no missing values. Fortunately, missing values presented in the PE log account for only 68 data points. For other types of logging, the missing values are substituted by the number -999.25. This study expects models to recognize -999.25 as missing values since this number is normally used in the petrophysical field symbolled for absent values (Thongsame, 2018).

4.2.2 Exploratory Data Analysis (EDA)

Exploratory data analysis (EDA) is an essential key for any research analysis to understand the overall trend and correlation of the data (Natrella, 2010). The well log data are examined for distribution, anomalies, and feature correlation via a variety of visualization techniques. This study aims to reconstruct well logs by existing well logs. Hence, well log data obtained as LAS files are converted to well log curves representation to visualize the data (Figure 4.3). Side-by-side boxplots are also used to see mean and log values distribution between multiple wells (Komorowski et al., 2016). Moreover, these boxplots can provide an insight into log anomalies if presented. Feature selection frequently falls into this step since EDA assists the researchers to find a pattern or correlation of input features as well.

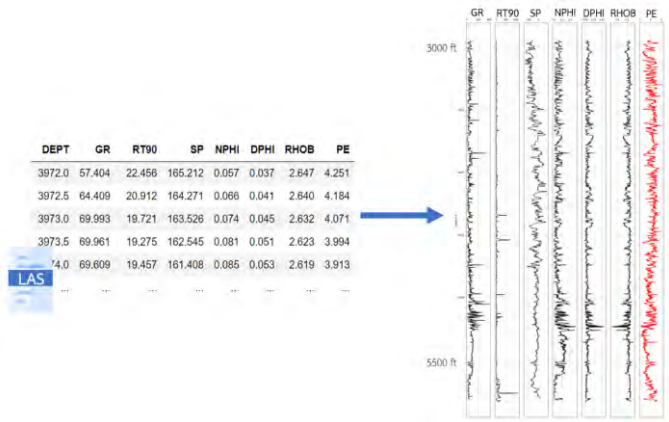


Figure 4.3 LAS files contain well log responses of each logging type in the form of numerical data at each depth. Converting them to log curves can provide an overall picture of well log data. PE log that this study aims to reconstruct is highlighted in red curve.

4.2.3 Data Transformation

There are three models applying in this study which are extreme gradient boosting (xGBoost), support vector regression (SVR), and artificial neural network (ANN). xGBoost, apart from the other two, is tolerant with different input feature scales because its algorithm is a rule-based system. The other two are considered sensitive to feature scales (Crone et al., 2006). Large measurement values can overwhelm the model performances. So that, to reduce the effects of large values, feature scaling is required. Regular ways to transform the data is min-max scaling and z-score shown in Equation (1) and Equation (2) respectively (Atomi, 2012.; Bisgin et al., 2018; Crone et al., 2006; Jayalakshmi and Santhakumaran, 2011; KumarSingh et al., 2015; Malaki et al., 2014). Min-max scaling changes data scale into a specific range e.g. [0,1]. While z-score is to transform your data that has a mean of 0 and a unit variance. This study applies Z-score for scaling well log features.

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \qquad \text{Equation (1)}$$

Where, x' s the min-max score, x is the original input data, x_{min} is the minimum value of the data, and x_{max} is the maximum value of the data.

$$x' = \frac{x - x_{mean}}{\sigma} \qquad \text{Equation (2)}$$

Where, x' in equation (2) is the standardized score, x is the original input data, x_{mean} Is the mean value of the data, and σ is the standard deviation of the data.

4.2.4 Data Partitioning

In general, the dataset will be divided into two main sets: a training data and a test data. However, overfitting is a key issue when trying to build a predictive model. Overfitted model is generally a complex model that perfectly explains the training data, even the noises. Resulting in good prediction for the training data but unable to predict test data or unseen data nicely (Berrar et al., 2013). The issue occurs due to the model lacking generalization property and the data containing high variance. To enhance the model generalization and reduce the data variance,

cross validation is thus applied (Thongsame, 2018). Cross-validation is a method to resampling data and assessing model generalization which k-fold cross-validation is often used in practice (Berrar, 2019). Well logging data was divided into three sets: a training set, a validation set, and a testing set. Data from the Kysar 1-1 well was used as the testing set while data from the other eleven wells were combined and resampled by 4-fold cross-validation technique: 70% for a training set and 20% for a validation set in each fold (Figure 4.4). The idea is that Kysar 1-1 well was excluded as an unseen data (shown in pink) and the rests are subsampled into 4 equal-size pieces (Berrar, 2019). The model is trained on the training set (shown in black), and optimized and evaluated on the validation set (shown in blue) for four iterations. In this method the validation set is partitioned in a way that there is no overlap of the data in each iteration.

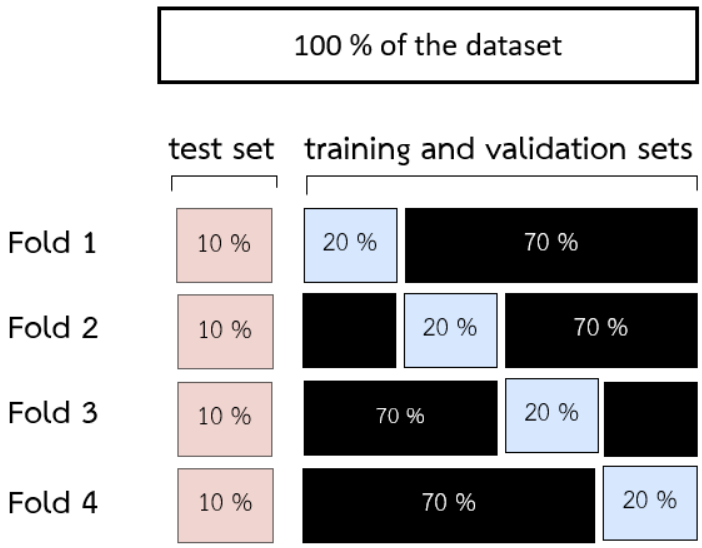


Figure 4.4 An example of 4-fold cross-validation where: pink represents testing sets, black represents training sets, and blue represents validation sets.

4.3 Well log reconstruction

To reconstruct well log responses, there are two main phases: training phase or model development phase, and model evaluation phase.

4.3.1 Model development phase

Extreme gradient boosting (xGBoost), support vector regression (SVR), and artificial neural network (ANN) were developed to reconstruct synthetic photoelectric (PE) logs. Each model was trained and optimized four times or folds corresponding to the 4-fold cross validation technique. Kysar 1-1 well was kept away from this phase. Other eleven wells were mixed and randomly subsampled into 4 equal subsets: three for training and one for validating. For the first fold, the model was trained on the training data which the algorithm aims to reduce the prediction errors on this dataset as much as possible shown in a black line on the error graph (Figure 4.5).

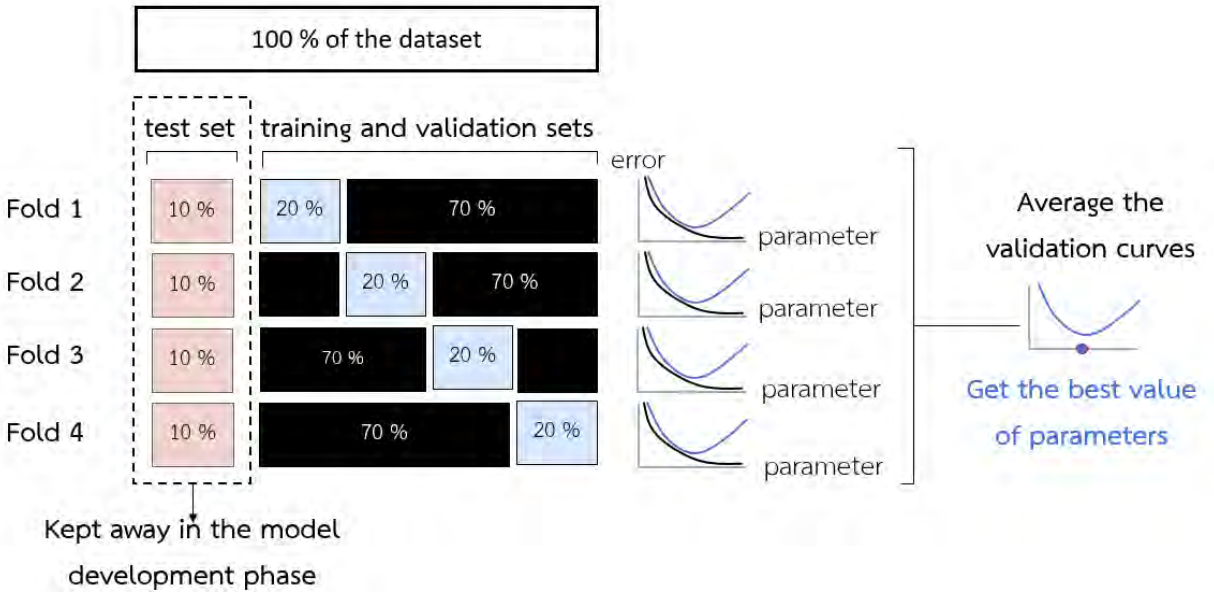


Figure 4.5 Model development phase where the model was trained and validated four times. Error graph is an output of the training and validating. The hyperparameter is plotted in different search ranges on the x-axis and prediction errors corresponding to the hyperparameter values plotted on the y-axis. The final cross-validation error graph is where the optimal hyperparameter range is selected. Note, black color stands for the training data related, blue stands for the validation data related, and pink stands for the test set.

After that, the model will be evaluated with the validation data where the prediction errors are more generalized than the training errors shown in a blue line on the error graph. Note that to plot an error graph y-axis is the occurring prediction error, and x-axis is the hyperparameter

of the model. *Learning rate* and *maximum depth*, for instance, are the hyperparameters of xGBoost. Likewise, for the second, third, and forth fold differ only in the data subsets. The cross-validation error graph is an average of all errors achieved on validation sets (Berrar, 2019). Since the x-axis of the cross-validation error graph is a model hyperparameter thus, this graph favors the researcher to locate an optimal range of that hyperparameter. This process is essential in improving model performance, so called hyperparameter tuning. Table 4.2 details the search range values of hyperparameters for each model. After optimal ranges were defined, grid search technique was applied to find the most optimized model.

4.3.2 Model architecture

4.3.2.1 xGBoost

This model is a tree-based algorithm thus the architecture is a tree-based architecture (if-then-else architecture)

4.3.2.2 SVR

An appropriate kernel function transforming input data into a higher dimension space is needed to be verified. Crone et al. (2006) study suggested that radial based function (RBF) is the most widely used kernel for regression problems. The RBF kernel is thus applied in this study as well.

4.3.2.3 ANN

According to Parapuram et al. (2018), a three layers neural network (2 hidden layers and one output layer) successfully predicts shear wave velocity with an adjusted R-square of 0.88. Onalo et al. (2018) also uses a three layers neural network to synthesize a compressional and shear transit time logs which root mean square error (RMSE) is 2.62 and 5.29 respectively. Therefore, this study adopted a three layers structure with backpropagation algorithm as an ANN architecture, both studies use backpropagation algorithm as well. In addition, dropout is also employed in the structure to reduce overfitting which is located on each of the fully connected (dense) layers before the output shows in Figure 4.6 (Hinton et al., 2012).

Model	Hyperparameter		Search range	References for the search ranges
XGBoost	Maximum depth		1 - 13	Thongsame et al. (2018)
	Learning rate		0.01 - 0.3	Thongsame et al. (2018)
SVR	Regularization (C)		0 - 200	Cherkassky et al. (2004); Thongsame et al. (2018)
	Epsilon		0.01 - 0.8	Cherkassky et al. (2004)
ANN	Number of nodes	1st hidden layer	1 - 64	Thongsame et al. (2018)
		2nd hidden layer	1 - 64	Thongsame et al. (2018)
	Dropout		0.05 - 0.4	Thongsame et al. (2018)

Table 4.2 Search range for hyperparameter tuning in each algorithm, two hyperparameters were adjusted in each algorithm. Since ANN has 2 hidden layers thus a number of nodes are needed to be tuned on both hidden layers.

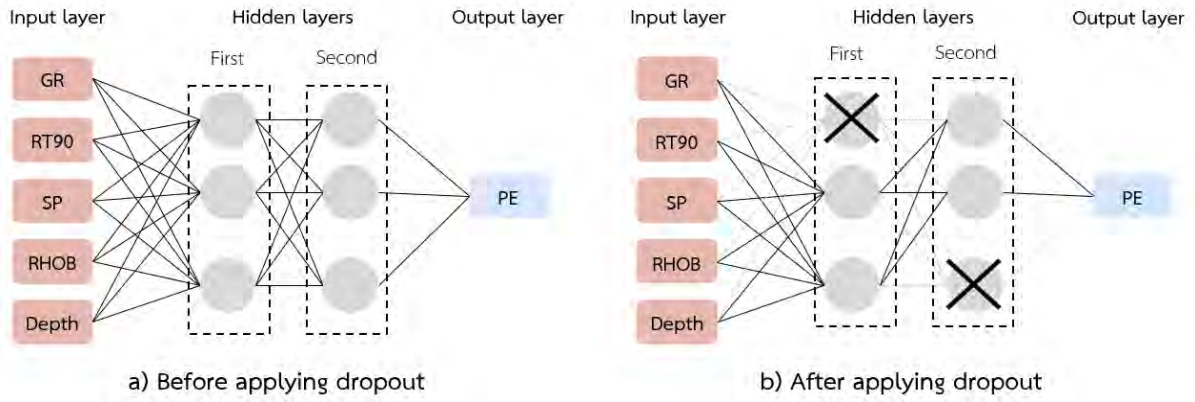


Figure 4.6 a) An example of a three-layered neural network, 2 hidden layers and an output layer before applying dropout, b) An example of the neural network after applying dropout. Dropout will set some nodes to zero to avoid overfitting. Note that input features are gamma ray (GR), deep resistivity (RT90), spontaneous potential (SP), bulk density (RHOB), and depth in feet and output is photoelectric (PE) log.

4.3.3 Model evaluation phase

The three best models, xGBoost, SVR, and ANN, already created from the previous phase. This phase is where the best models first encounter with the test set or unseen data that was kept away from the development phase. The test data, Kysar 1-1 well, was given to the model and the synthetic PE log, predicted PE, was served as an output. The predicted PE was evaluated by comparing with the actual PE responses. Mean square error (MSE) and R-square were used as evaluation metrics (Figure 4.7).

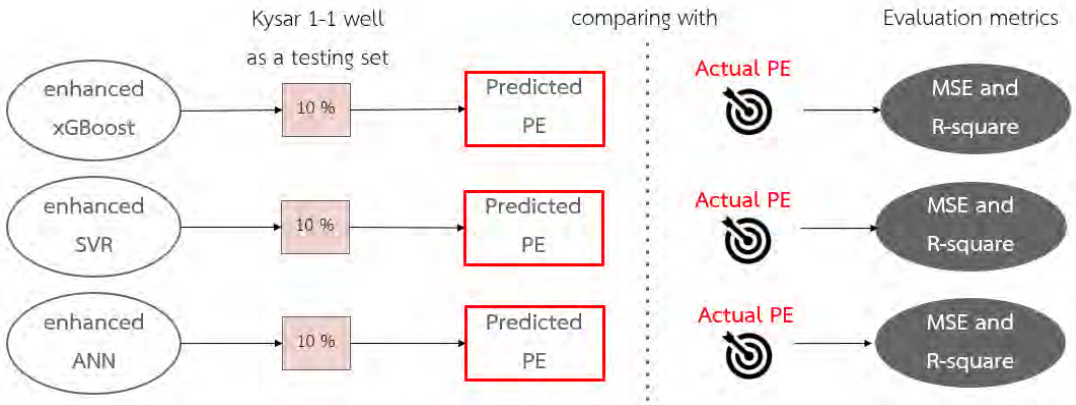


Figure 4.7 Model evaluation phase

4.4 Model comparison

The performance of three models were compared in this step. Datasets used to train, validate, and test the model are identical for every model. Independent variables are different model algorithms and dependent variables are performances of each algorithm. To determine model performances, MSE and R-square are adapted. MSE is a method to measure the differences in values between the predicted and the actual values. Whereas, R-square or coefficient of determination is generally used to show how much variance can be explained by the model (Parapuram et al., 2018).

Chapter 5 Results

5.1 Exploratory Data Analysis

Twelve wells were collected with a total of 51,385 data points. The depth of wells ranges from 300 ft - 5800 ft. A common 6 well logging types were used in this study: gamma ray (GR), photoelectric factor (PE), deep resistivity (RT90), spontaneous potential (SP), density porosity (DPHI), and bulk density (RHOB). As mentioned before that well logging operation requires intensive labors and considerably high costs, hence some wells may contain missing data (Figure 5.1). This study aims to synthesize PE log resulting in removing data points at the depth where the PE log values have vanished. 51,338 is the final total amount of prepared data with only 3 percent of missing values. The absent values are replaced by -999.25, a number representing a missing value in the petrophysical field. This number is used expecting models to recognize it as missing values (Thongsame, 2018). The range value of each well logging type comparing between 12 wells is illustrated in Figure 5.2. Average log values, in most wells, of each logging type are alike, yet only HCU 2220-B well differentiates from the rest. 'PE' boxplot, Figure 5.2a, shows that the average PE log responses of the HCU 2220-B well is lower than the other wells, which approximately equal to 1.6 - 2. Thus, according to Glover (2012), the average value in this well is a reflection of quartz mineral (sandstones). Meanwhile, the average value of the other wells is roughly a reflection of dolomite and calcite (evaporitic rocks). For 'RT90' boxplot, Figure 5.2b, the average value of the HCU 2220-B well is the same as others, but there are still some peculiar outstanding values needed for further investigation. The data of this well is collected at a shallow distinct depth range from most of the wells (Table 4.1). Besides after overseeing the well log responses in Figure 5.1, it suggests that the notable deep resistivity value spikes at a depth around 2,250 ft., but corresponding with the rapid changes in other logging types at the same depth as well. Hence, these values are not anomalies.

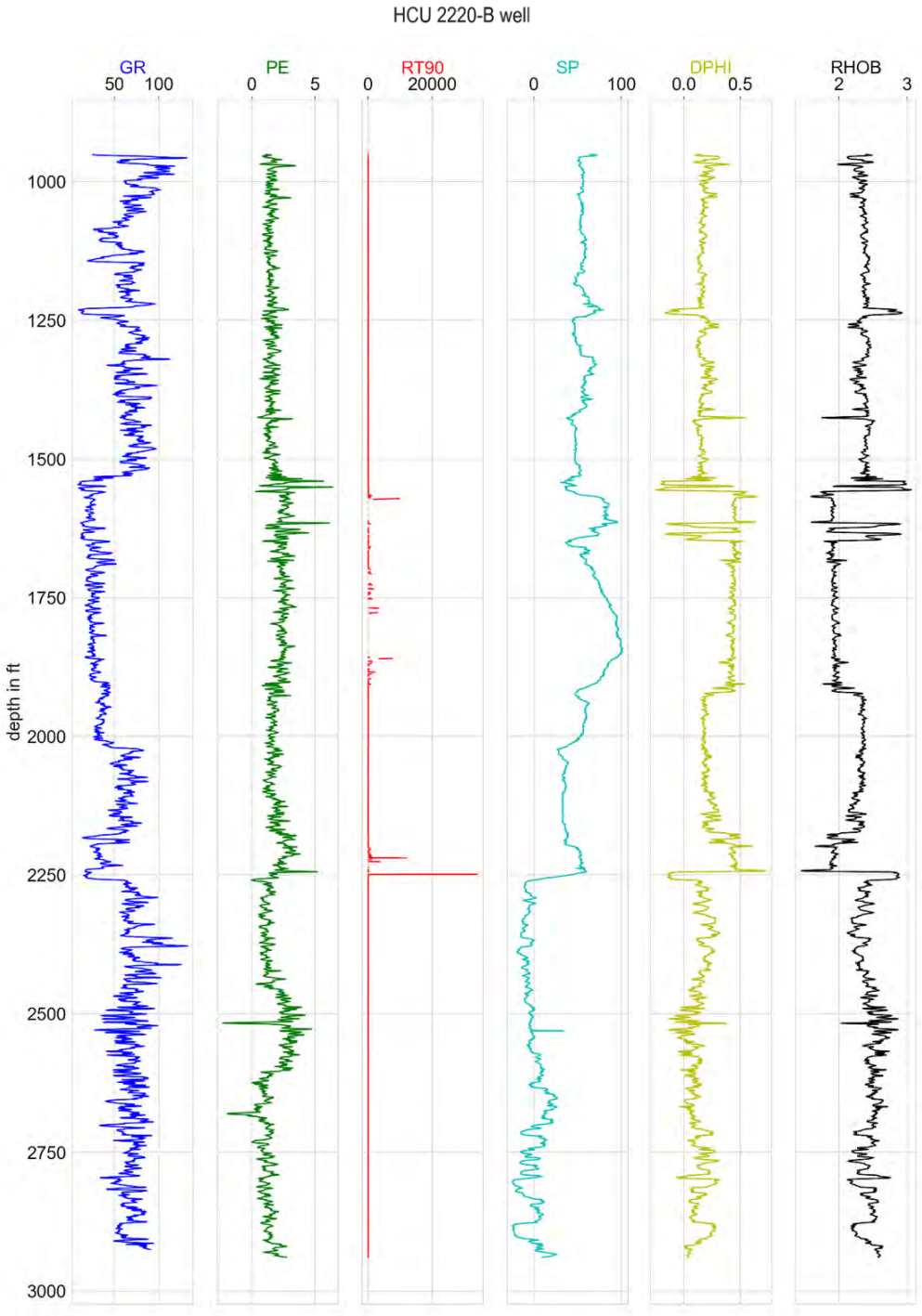


Figure 5.1 An example of missing data in RT90 log presented in HCU 2220-B well log responses where: blue color for gamma ray (GR) log, green for photoelectric (PE) log, cyan blue for deep resistivity (RT90) log, yellow for density porosity (DPHI) log, and black for bulk density (RHOB). Notice missing data points at different depths in RT90 log which are evidenced by discontinuous responses.

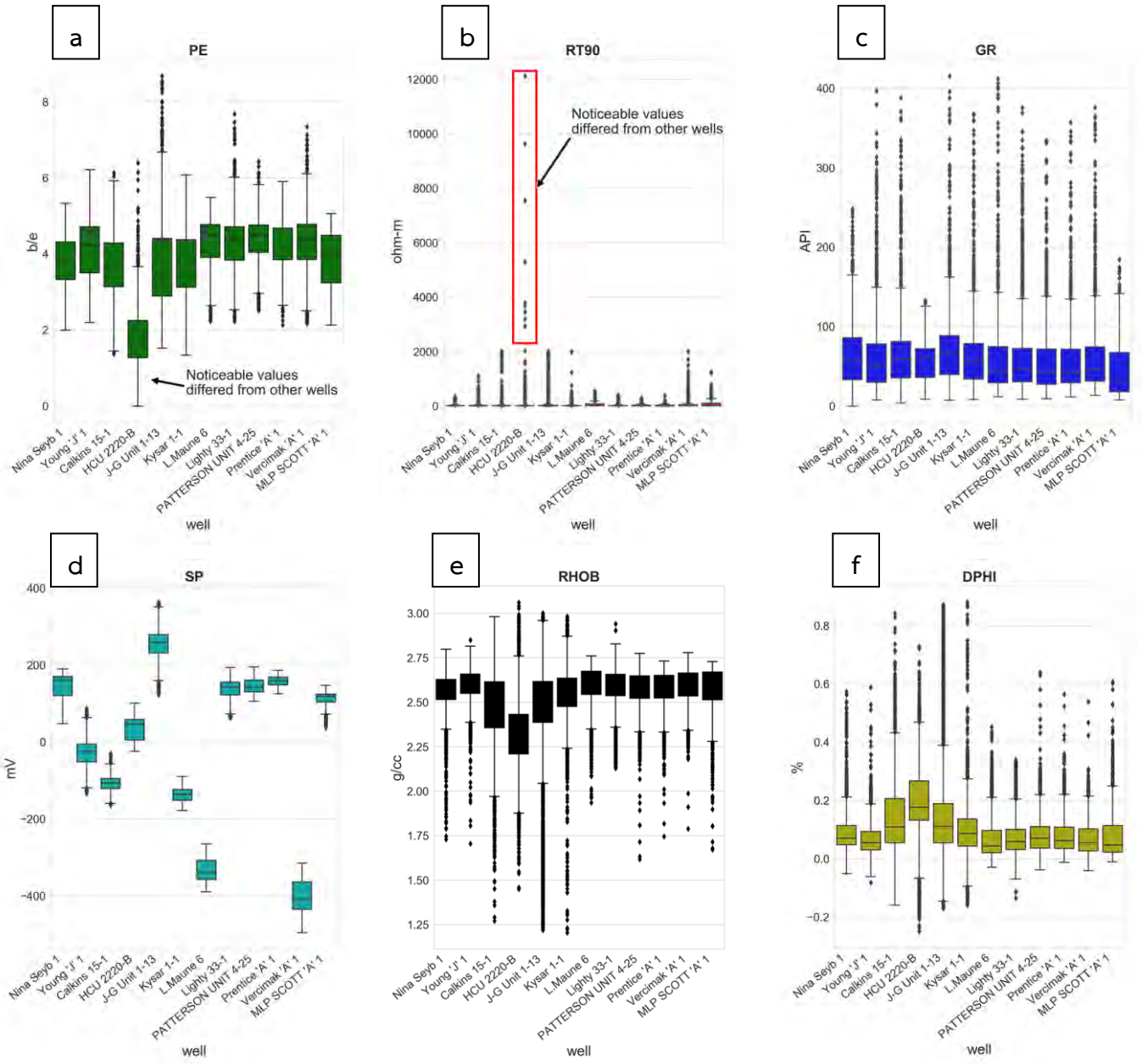


Figure 5.2 Boxplots showing range value of each well logging type comparing between 12 wells. a) boxplots for photoelectric factor log responses, b) deep resistivity log responses, c) gamma ray log responses, d) spontaneous potential log responses, e) bulk density log responses, and f) density porosity log responses.

However, only 4 features of well logging (GR, RT90, SP, RHOB) and depth (in feet) were used as input data in the model to reconstruct synthetic photoelectric (PE) log. Density porosity (DPHI) was excluded. DPHI and RHOB are obtained from the same density log, which measures gamma rays left from Compton scattering and only differs in the calculation method. Based on obtained well logging data, DPHI and RHOB data are highly correlated with Pearson (linear) correlation of -1 (minus means higher in DPHI, lower in RHOB) suggesting that these two features simply can represent each other (Figure 5.3). Thus RHOB, which has a higher Pearson correlation with PE, is preserved and DPHI is excluded.

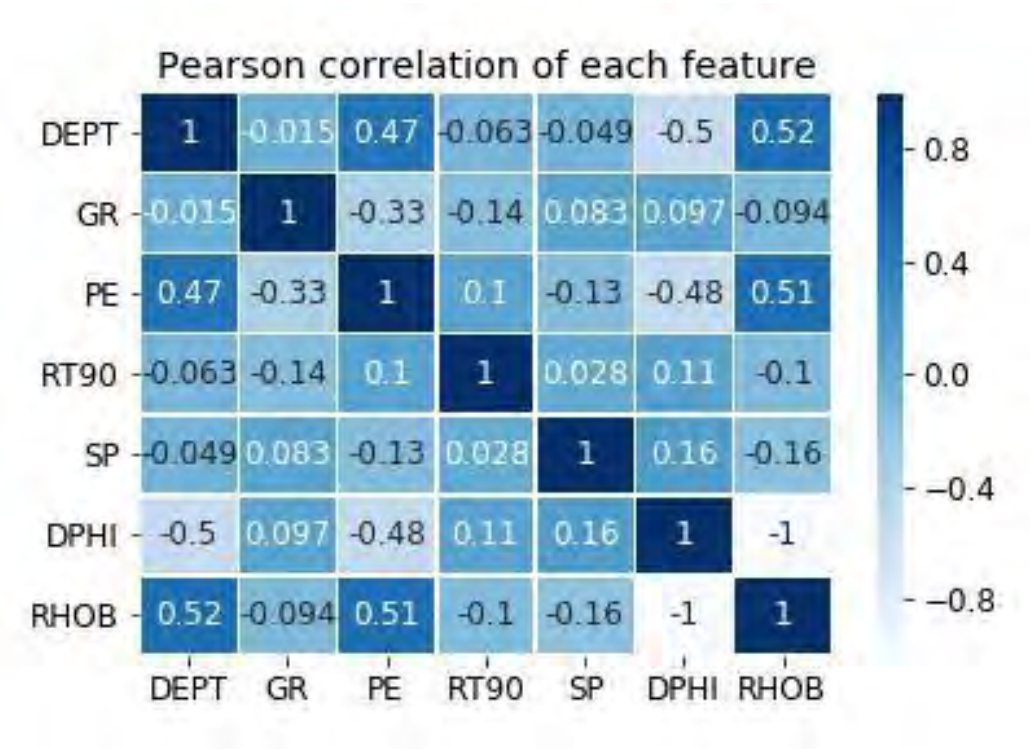


Figure 5.3 Heatmap showing Pearson correlation between each feature. DPHI and RHOB logging responses are highly correlated.

5.2 Model Performance

Well logging data was divided into three sets; 70% for training set, 20% for validation set, and 10% for testing set. Data from the Kysar 1-1 well was used as the testing set while data from the other eleven wells were combined and used as training and validation sets. The combined data was then divided into two datasets: the training set, which accounts for 36,000 data points, and the validation set, which is used for hyperparameter tuning. A model was first trained on the

training set and parameters were optimized by 4-fold cross-validation using the validation set. This technique favors the model to reduce the variance of the dataset (Thongsame, 2018). After validation, the optimized model was evaluated by the testing set using error functions such as mean square error and R-square methods. Mean square error, MSE, is one of the loss function algorithms attempting to reduce in the training phase. MSE is thus used to select the best model in this study indicated by the lowest MSE. Nevertheless, MSE values can range in a wide range depending on the unit range of measurement. Akinnikawe et al. (2018) reconstructs the PE logs and the unconfined compressive strength (UCS) logs from conventional well logs by various means of machine learning. Results show that random forest, the best model, can predict the PE logs with MSE of 0.33 while artificial neural networks can predict UCS logs with MSE of 320.2. Because the unit range of PE logs varies between 1.7 - 5.1 b/e and varies between 1,700 - 40,000 psi for UCS logs. R-square, consequently, is a helpful tool to compare results from different logging types. R-square is a closeness measurement between the predicted data and the actual data (Parapuram et al., 2018). The results from Kysar 1-1 well were shown since it Is the testing data or the unseen data which models have never encountered. While data from the other eleven wells were already used to train and validate the models, thus the predicted result of those wells are not generalized.

The first model to mention is extreme gradient boosting (XGBoost) since this model has the lowest mean squared error, MSE, of 0.139 (or R-square of 0.75) determining it to be the best model. For support vector regression (SVR), which is a type of support vector machine handled with regression problems, has an MSE of 0.154 (or R-square of 0.71). Artificial neural networks surprisingly generated the highest MSE, the worst model, of 0.197 (or R-square of 0.65). Noteworthy that the input data for SVR and ANN is transformed by z-score procedure before training the model. MSE, R-square results, and relation between predicted PE and acutial PE of each model are illustrated in Figure 5.4. The actual PE and the predicted PE from each model is plotted as well log curves in Figure 5.5.

To inform that, the ideal image of scatter plot between the actual PE on the x-axis and the predicted PE on the y-axis is a linear trend of the data points with an R-square equals to 1. In Figure 5.4b. the data points in the scatter plot of xGBoost are clustered in a linear trend the

scatter plot of xGBoost result suggests that the predicted PE log values are covered from 2 to 5 b/e and the model predicts slightly poorly when the actual PE is less than 2 and more than 5. For the scatter plot of SVR result, Figure 5.4c, a few predicted results can go high up to around 6.5, this may be exemplified at depth around 4,150 ft. in Figure 5.5b. However, the data points are still well scattered in a linear trend. The last model is ANN, Figure 5.4d, the data points are not scattered in a linear trend, moreover, the scatter plot yields that ANN is unable to provide outputs less than 3 b/e and higher than 5 b/e. Therefore, the predicted result of this model is slightly left behind because the actual PE log values ranging from 1 to 6 b/e are constrained to be predicted in a range of only 3 to 5 b/e. Figure 5.5 is used to see the predicted PE responses along with the actual PE responses where the solid red lines are the prediction of each model, and the dashed blue lines are the actual PE responses. There are two noticeable different sections; the first section is where the average value of PE is around 2.8 b/e, can be interpreted as shaly sandstone, located at depth less than 2,750 ft. Another succession is where the average value of PE is 4.1 b/e, interpreted to be evaporitic rocks, detected at depth more than 2,750 ft. For the first section, xGBoost and SVR perform as good as each other (Figure 5.5a, b), while ANN performs unsuccessfully (Figure 5.5c) with an outstanding line at a predicted value of 3 b/e. On the other hand, the second section is where all three models can perform similarly well.

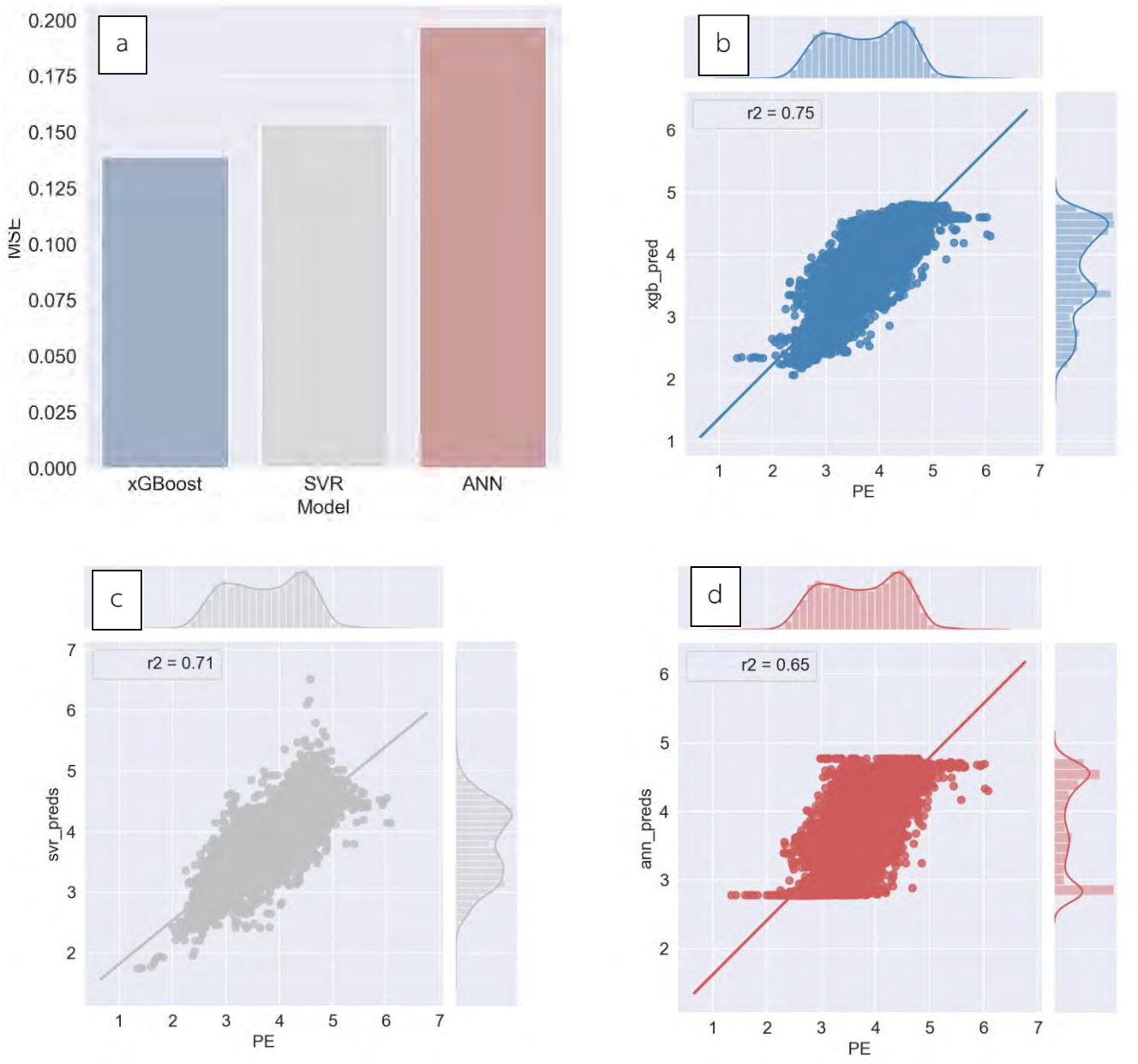


Figure 5.4 a) Mean square error of each model performance: lowest in xGBoost means the model best synthesizes the PE log. b) A jointplot showing R-squared result of the xGBoost model where: actual PE is plotted on x-axis and its distribution is projected upward and plotted on the top. The predicted PE is plotted on y-axis and its distribution is projected rightward and plotted on the right. This figure configuration is applied in all joint plots. c) A jointplot showing R-square result of SVR, and d) a jointplot showing R-square result of ANN.

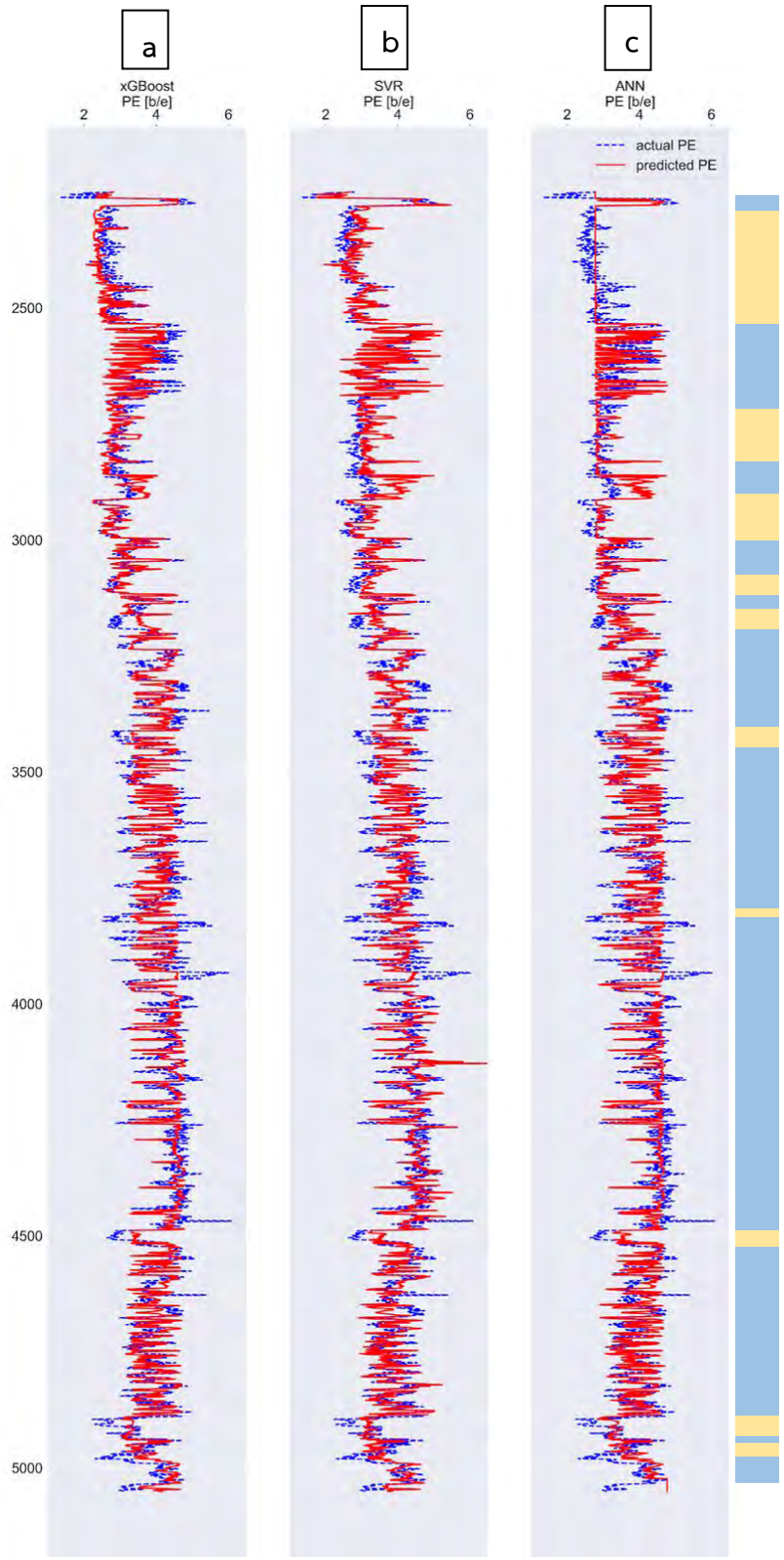


Figure 5.5 Actual and predicted PE along with depth (feet) of different algorithms for Kysar 1-1 well of a) xGBoost model, b) SVR model, and c) ANN model. Solid red lines represent the predicted PE, the dashed blue lines are the actual PE, yellow boxes on the right indicated to be shaly sandstone, and blue boxes are evaporitic rocks.

5.3 Effect of hyperparameter tuning

Before getting the final MSE or R-square, 4-fold cross-validation was applied to find the best hyperparameters in each model. Since there is no model that can predict any kind of dataset accurately without an adjustment. Hyperparameter tuning can find the optimal values in the model for better performance and generalization.

For xGBoost, *learning rate* and *maximum depth* parameters were optimized. *Learning rate* controls how fast a tree, added to xGBoost sequentially, learns to correct the residual errors from the previous tree. *Maximum depth* controls the maximum depth of each tree feeded to the xGBoost. Table 5.1 shows the search range, optimal range, and the most optimized values of hyperparameters for each model. The optimal range of *learning rate* and *maximum depth* were selected by the graph showing in Figure 5.6. The optimal range for *maximum depth* is between 3 and 5. Although the number of *maximum depths* increases, the MSE does not decrease significantly. For *learning rate*, the optimal range is obviously between 0.025 and 0.05 because of the dogleg style of the graph. After optimal ranges were defined, grid search technique was applied to find the most optimized model which *maximum depth* is 4, and *learning rate* is 0.045.

Model	Hyperparameter		Search range	Optimal range	Optimized value
XGBoost	Maximum depth		1 - 13	3 - 5	4
	Learning rate		0.01 - 0.3	0.025 - 0.05	0.045
SVR	Regularization (C)		0 - 200	40 - 60	50
	Epsilon		0.01 - 0.8	0.4 - 0.6	0.4
ANN	Number of nodes	1st hidden layer	1 - 64	8 - 16	13
		2nd hidden layer	1 - 64	3 - 16	15
	Dropout		0.05 - 0.4	0.05 - 0.1	0.05

Table 5.1 The search range, optimal range and the most optimized value obtained from the grid search technique for every model. Since ANN has 2 hidden layers thus a number of nodes are needed to be tuned on both hidden layers.

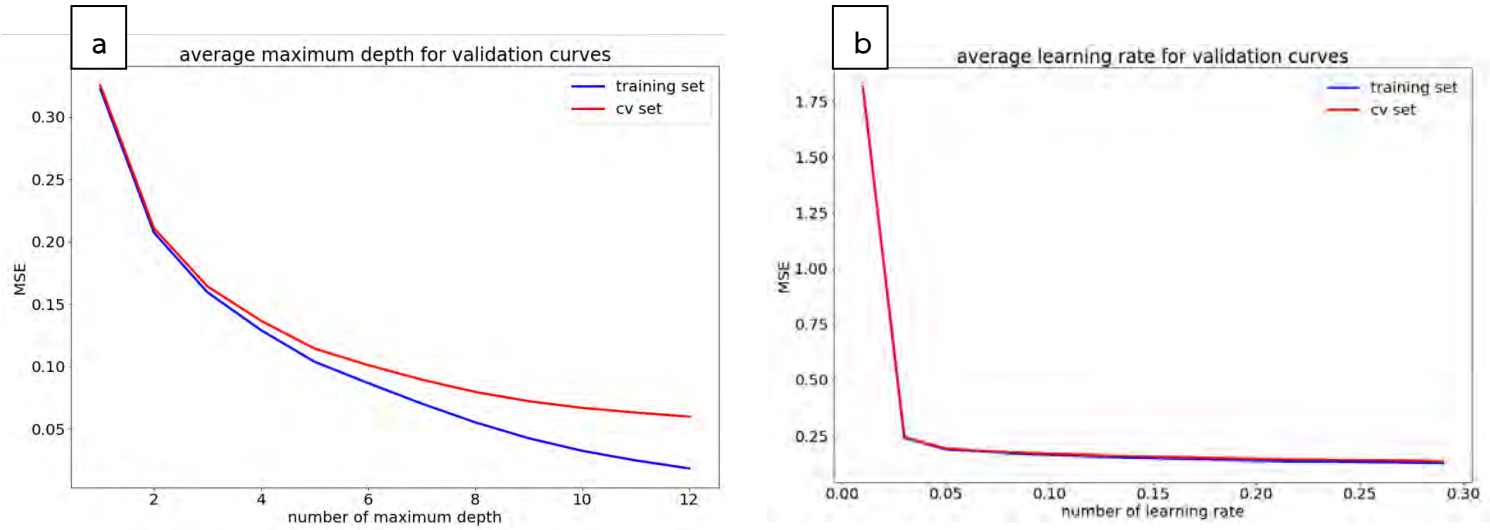


Figure 5.6 Line graphs represent the effect of tuning parameters to MSE in XGBoost a) the effect of *maximum depth*, amd b) the effect of *learning rate*

For SVR, *C* and *epsilon* were tuned. *C* is a regularization term indication of how much the model can tolerate misclassifying of each training example. Higher *C* means the model cannot tolerate any misclassified data leading to overfitting. In contrast, lower *C* can tolerate many misclassified data leading to underfitting. To optimize generalization of the SVR model, errors within a certain distance, epsilon, are ignored. Figure 5.7 shows that the optimal range for *C* is between 10 and 50 (Figure 5.7a) since *C* of more than around 50 is slowly decreasing in MSE. *Epsilon* both in the training set and validation set reach their bottoms between 0.4 to 0.5. Furthermore, *epsilon* of more than 0.5 will lead to the rising of MSE rapidly.

Number of nodes in the hidden layers both first and second layers and the *dropout* are adjusted in ANN. Rate of *dropout* is used to prevent overfitting by setting a chance to some number of layer outputs, nodes, to zero (or dropped out). The optimal range for the *number of nodes* in the first hidden layer is 8 to 16 nodes (Figure 5.8a), for the *number of nodes* in the second hidden layer is 3 to 16 nodes (Figure 5.8b), and for the rate of *dropout* is 0.05 to 0.1 (Figure 5.8c).

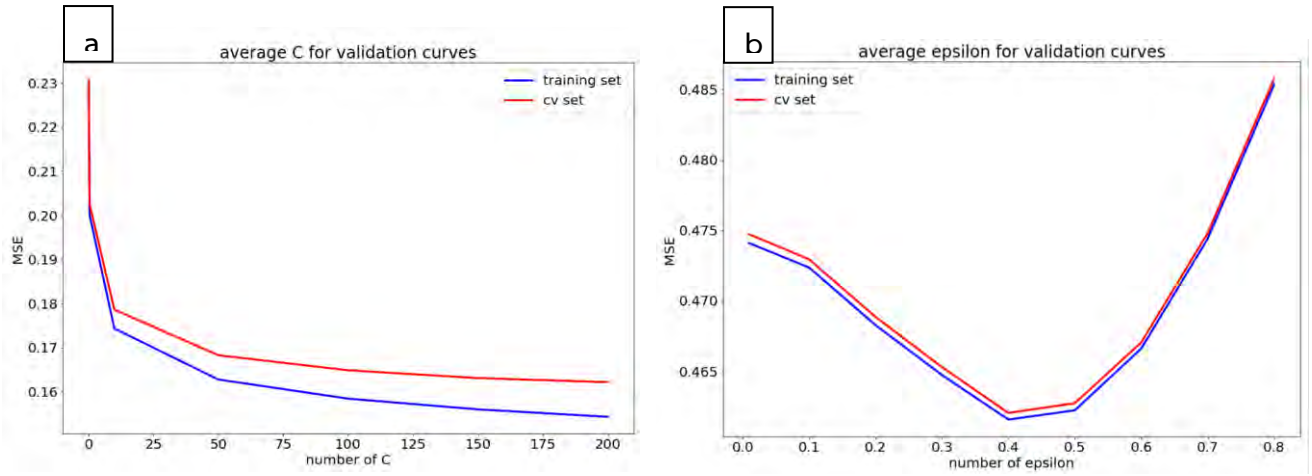


Figure 5.7 Line graphs represent the effect of tuning parameters to MSE in SVR a) the effect of *C*, and b) the effect of *epsilon*

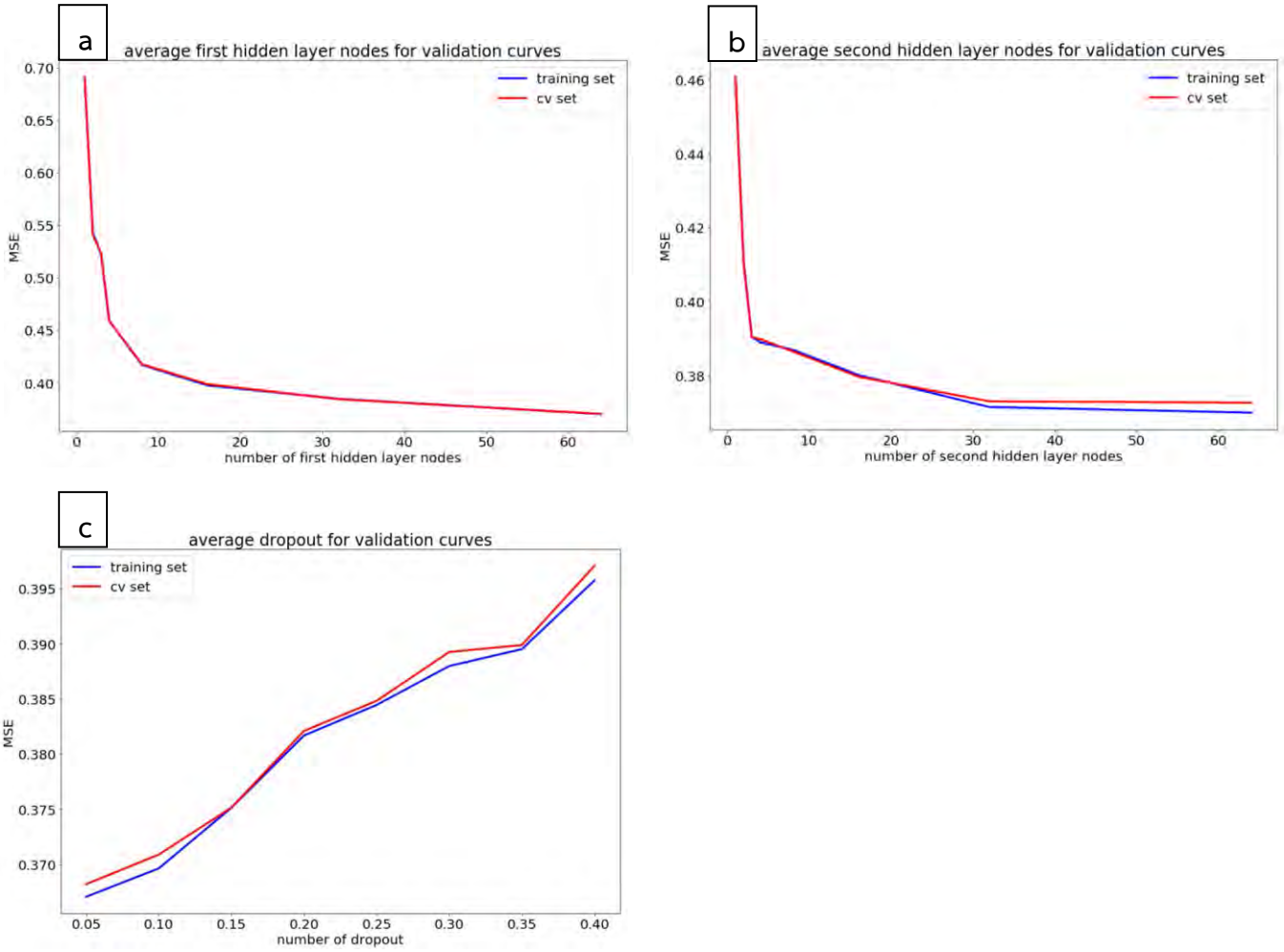


Figure 5.8 Line graphs represent the effect of tuning parameters to MSE in ANN a) the effect of *number of nodes* in the first hidden layer, b) the effect of *number of nodes* in the second hidden layer, and c) the effect of *dropout* rate

Chapter 6 Discussion and Conclusions

6.1 Discussion

6.1.1 Model performance

Extreme gradient boosting (xGBoost), support vector regression (SVR), and artificial neural network (ANN) are utilized to predict photoelectric (PE) logs from convention well logs. Gamma ray (GR), spontaneous potential (SP), deep resistivity (RT90), and bulk density (RHOB) along with depth in feet are the input data. Moreover, in the training phase, z-score and min-max scaling were used to transform the input data to reduce the effect of scaling in SVR and ANN. Models with min-max scaling method perform worst than z-score method. Since min-max scaling will only change the range of data to specific range such as [0,1] or [-1,1] thus it is significantly sensitive to outliers. in this case, outliers mean a wide range of log values such as the range of GR which is distributed between 0 to 415 API with a mean of 60.2 API (right-skewed distribution shown in Figure 6.1). Meanwhile, z-score method provides a zero mean and a unit variance which its advantage is that it reduces the effect of outliers (Atomi, 2012). XGBoost provides the least MSE of 0.139 (R-square of 0.75) which is considered to be the best model to synthesize PE log in this study. SVR is the second best with MSE of 0.154 (R-square of 0.71) and the worst model is ANN with MSE of 0.197 (R-square of 0.65).

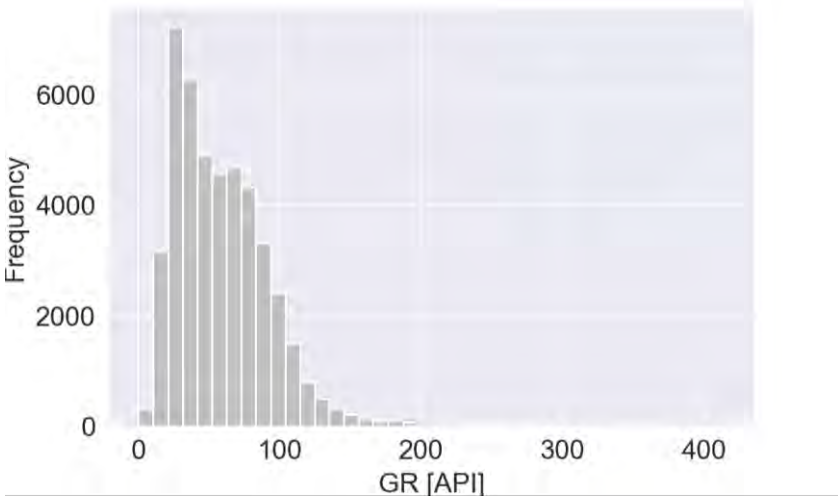


Figure 6.1 Gamma ray (GR) distribution plot showing right-skewed of the data with a wide range of values (0-415 API).

Comparing results between three algorithms

Figure 6.2 illustrates the major lithology that was obtained from twelve wells located in the Hugoton embayment, Kansas. From the figure, it can be seen that evaporitic rocks, PE average is around 4.1 b/e, are the most common data while there is only 10% of shaly sandstone data, PE average is around 2.5 b/e. The unequal amount of lithology or data is also known as imbalanced data which is the common problem encountered in real-world dataset. Imbalanced in lithology results in imbalanced PE values as well, since PE value can directly represent the main mineral component of each rock type. For example, quartz which is the main mineral for sandstone has PE value about 1.8 b/e, shaly sand for shaly sandstone has PE value around 2.7, shale has PE value around 3.42 b/e, and PE value more than 4 are indication for evaporitic rocks (Glover, 2012). The results suggest that the predicted PE logs from xGBoost and SVR can align along with the actual PE logs both in evaporitic succession and shaly sandstone succession. ANN performs well in evaporitic succession but not in the shaly sandstone succession due to the influence of imbalanced data problem. Ustuner et al. (2016) performed a study on RapidEye imagery classification with balanced and imbalanced data with three machine learning algorithms: support vector machine (SVM), maximum likelihood (ML), and artificial neural network (ANN). The study informs that SVM outperforms ML and ANN. The overall classification accuracy of ANN reduced 3-5% with imbalanced data. In contrast SVM is a kernel-based classification algorithm that can tolerate imbalanced data by using a kernel function (Ustuner et al., 2016).

Comparing results with other studies

Studies performed log reconstruction compared with this study shown in Table 6.1. Some of the mentioned studies predict the PE logs similar to this study thus MSE and R-square can both be used to compare model performances directly. Only one study predicted different kinds of logging types thus only R-square can be used to compare the result. First, consider the results from Akinnikawe et al. (2018) and this study, they assure that the tree-based algorithm can synthesize PE logs from conventional logs with satisfying results. Because xGBoost, a tree-based algorithm, is the best model in this study to reconstruct the PE log and the first three models in Akinnikawe et al. (2018) are the tree-based algorithms as well.

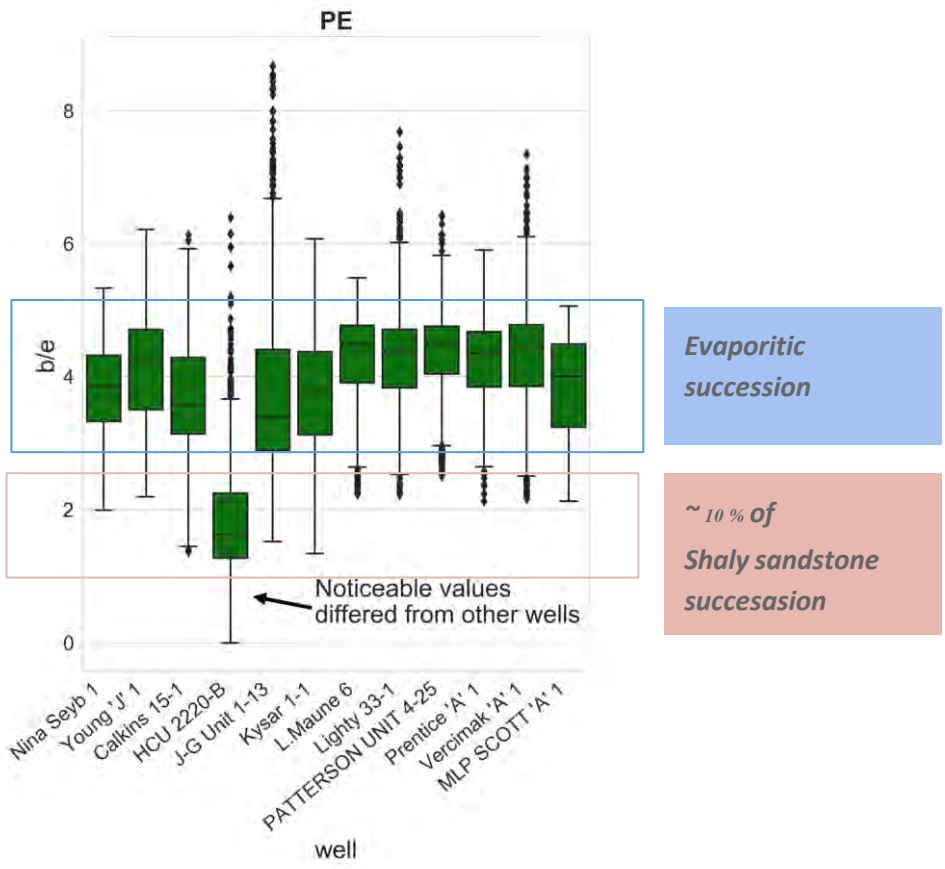


Figure 6.2 Boxplots of PE value distribution comparing between 12 wells. Evaporitic rocks are the main lithology in this study area with 10% of shaly sandstone.

Tree-based algorithms have an advantage in working as an if-then-else system resembling human decisions. In addition, xGBoost has parameters to deal with imbalanced data (xgboost developers, 2016) which has been proven in many researches (Stanley et al., 2020; Thongsame, 2018; Sun et al, 2019).

Considering MSE between this study and Akinnikawe et al. (2018) study, there are not many differences indicating that the errors in this study are acceptable. This study, even though, has a lower MSE but that because of the advantage of this study. The dataset utilized in Akinnikawe et al. (2018) is acquired from over 100 well log files which means more in lithologies and heterogeneity of the data while this study uses 12 well log files. Asoodeh and Shadizadeh (2015) study which works on committee neural networks (CNN) has the lowest MSE result. CNN result is obtained from the combination of three different architectures of neural networks: radial basis neural networks (RBNN), bayesian regulation neural networks (BRNN), and generalized regression neural network (GRNN). Each neural network results then combined with the favor of

a genetic algorithm-pattern search (GA-PS) thus the final predicted PE of CNN model is more generalized. As can be known, SVR is not used to generate the PE log, but Wong et al. (2005) used SVR and ANN to generate porosity data instead. The input data for generating porosity data is a suite of well logging as well. The result shows that SVM (MSE of 23.71) outperforms the ANN (MSE of 66.68). Since prediction accuracy of ANN depends much on the generalization ability and real-world well log data is too noisy for ANN (Wong et al. 2005). Moreover, SVR has a regularization term to tolerate noise in the data (Huang et al., 2008). Hence, Support vector machines or support vector regression is a powerful tool because it is a kernel-based algorithm that can perform well in both imbalanced data and noisy data. MSE of this study and Wong et al. (2005) can not be compared directly because different kind of logging types.

Articles	Input Data	Output Data	Model	R-square	MSE
This study	Depth, GR, RT90, SP, RHOB	PE	- xGBoost	0.75	0.139
			- SVR	0.71	0.154
			- ANN	0.65	0.197
Akinnikawe et al. (2018)	GR, RHOB, NPHI, RT90, NPHI, Vsh, and differences between NPHI and DPHI (new feature)	PE	- Random forest (RF)	-	0.33
			- Gradient boosting (GB)	-	0.348
			- Decision tree	-	0.349
			- ANN	-	0.38
			- Linear regression	-	0.395
Asoodeh and Shadizadeh (2015)	DT, RHOB, NPHI	PE	- Committee neural network (CNN)	0.879	0.02
Wong et al. (2005)	DT, RHOB, RT90, RT60, RT30, PE, GR, NPHI	Porosity	- SVM	-	23.71
			- ANN	-	66.68

Table 6.1 Result comparison between this study and others studies

6.1.2 Recommendation

Imbalanced data is a problem encountered in this study thus future works can increase the amount of minor class by collecting more data or using upsampling methods. For major classes where the amount of data can be reduced by downsampling method. So that, the PE distribution covers all lithologies. Feature engineering, moreover, can be proposed to improve the model performances. In addition, other models such as concurrent neural network (CNN) and recurrent neural network (RNN) can be applied to this dataset

6.2 Conclusions

This study aims to reconstruct synthetic PE logs with distinct machine learning algorithms: extreme gradient boosting (xGBoost), support vector regression (SVR), and artificial neural network (ANN). 12 well log data is obtained from the Hugoton Embayment which is a shelf extension of Anadarko basin in Kansas, USA. Kysar 1-1 well was kept as a test set and other wells were subdivided into a training set and a validation set. 4-fold cross-validation was used as a technique to generate different four training sets and four validation sets in order to increase model generalization and avoid overfitting. The models were built with hyperparameter tuning by the validation set and evaluated four times on different datasets. The final MSE and R-square, evaluation metrics in this study, are the average of four results.

From the exploratory data analysis, it can be concluded that 12 well log datasets are not balanced in lithologies. Evaporitic rocks are major rocks with smaller amounts of clastic rocks which is shaly sandstone. This can cause a problem in some machine learning algorithms.

All models can properly synthesize PE log since the occurring mean square error (MSE) is distributed in small ranges from 0.139 to 0.197. Even though, the data used in this study is imbalanced and models in this study are simple, but the results and MSE are considered acceptable thus the models are effective in synthesizing PE log

xGBoost becomes the best model to reconstruct the PE log in this study because of its ability to work as a human. In other words, xGBoost is a rule-based algorithm or if-then-else system resembling human decisions. xGboost also has parameters to deal with imbalanced data. MSE of xGBoost is 0.139 and R-square is 0.75 considered as a satisfying result.

xGBoost has the ability to select its feature importance, ordering from the most important, top, to the least important, bottom, respectively.

- Depth
- Gamma ray
- Spontaneous potential
- Deep resistivity
- Bulk density

SVR is the second-best model with 0.194 and 0.71 for MSE and R-square respectively. The result suggests that even though it does not work as a human decision, with a kernel-based algorithm, its kernel function can map lower-dimensional data to be higher dimension data. Two classes of data that can not separate linearly can be projected to non-linear separable data through a kernel function. Nevertheless, the model is robust, constant and effective under balanced and imbalanced dataset.

The last model is ANN which has the highest MSE of 0.197 and lowest R-square of 0.65. ANN performs worst since its sensitivity to the generalization of the data or the imbalanced data. If the data is imbalanced the model will try to learn only the major distribution of the data. Resulting in performing well with the major distribution of the data (evaporitic rocks with PE around 4 - 5, in this case), but poor on the minor distribution (lithologies with PE less than 4 and more than 6).

List of references

Aha, D.W. and Breslow, L.A. 1998. Comparing Simplification Procedures for Decision Trees on Economics Classiication Task. Naval Research Laboratory, Washington, DC, pp. 19.

Akhundi, H., Ghafoori, M. and Lashkaripour, G.-R. 2014. Prediction of shear wave velocity using artificial neural network technique, multiple regression and petrophysical data: a case study in Asmari reservoir (SW Iran). Open J. Geol. 04, 303–313.

Akinnikawe, O., Lyne, S. and Roberts, J. 2018. Synthetic Well Log Generation Using Machine Learning Techniques, in: Proceedings of the 6th Unconventional Resources Technology Conference. Presented at the Unconventional Resources Technology Conference, American Association of Petroleum Geologists, Houston, Texas, USA.

Asoodeh, M. and Shadizadeh, S.R. 2015. The Prediction of Photoelectric Factor, Formation True Resistivity, and Formation Water Saturation from Petrophysical Well Log Data: A Committee Neural Network Approach. Energy Sources, Part A: Recovery, Utilization, and Environmental Effects 37, 557–566.

Atomi, W.H. 2012. The effect of data pre-processing on the performance of Artificial Neural Networks Techniques for Classification problems 42.

Basak, D., Pal, S. and Patranabis, D.C. 2007. Support Vector Regression. Neural Information Processing 11, 23.

Berrar, D. and Dubitzky, W. 2013. Overfitting, in: W. Dubitzky, O. Wolkenhauer, K.-H. Cho, H. Yokota (Eds.), Encyclopedia of Systems Biology, Springer, 1617–1619.

Berrar, D. 2019. Cross-Validation, in: Encyclopedia of Bioinformatics and Computational Biology. Elsevier, 542–545.

Biau, G. and Cadre, B. 2017. Optimization by gradient boosting. arXiv:1707.05023 [cs, math, stat].

Bisgin, H., Bera, T., Ding, H., Semey, H.G., Wu, L., Liu, Z., Barnes, A.E., Langley, D.A., Pava-Ripoll, M., Vyas, H.J., Tong, W. and Xu, J. 2018. Comparing SVM and ANN based Machine Learning Methods for Species Identification of Food Contaminating Beetles. Sci Rep 8, 6532.

Buhulaigah, A., Al-Mashhad, A.S., Al-Arifi, S.A., Al-Kadem, M.S. and Al-Dabbous, M.S. 2017. Multilateral wells evaluation utilizing artificial intelligence. SPE Middle East Oil Gas Show Conf.

Burke, K. and Dewey, J.F. 1973, Plume-generated triple junctions; key indicators in applying plate tectonics to older rocks: Journal of Geology, v. 81, 406-433.

Cherkassky, V. and Ma, Y. 2004. Practical selection of SVM parameters and noise estimation for SVM regression. Neural Networks 17, 113–126.

Chitsazan, N., Nadiri, A.A. and Tsai, F.T.C. 2015. Prediction and structural uncertainty analyses of artificial neural networks using hierarchical Bayesian model averaging. J. Hydrol. 528, 52–62.

Chopra, P., Papp, É. and Gibson, D. 2005. Geophysical well logging.

Crone, S.F., Guajardo, J. and Weber, R. 2006. The Impact of Preprocessing on Support Vector Regression and Neural Networks in Time Series Prediction 6.

Demolli, H., Dokuz, A.S., Ecemis, A. and Gokcek, M. 2019. Wind power forecasting based on daily wind speed data using machine learning algorithms. Energy Conversion and Management 198, 111823.

Denison, R. E., Lidiak, E. G., Bickford, M. E. and Kisvarsanyi, E. B. 1984. Geology and geochronology of Precambrian rocks in the central interior region of the United States: U.S. Geological Survey Professional Paper 1241-C, 20 p.

Dubois, M.K., Bohling, G.C. and Chakrabarti, S. 2007. Comparison of four approaches to a rock facies classification problem. Computers & Geosciences 33, 599–617.

Giao, P.H. and Chung, N.H. 2017. A Case Study on Integrated Petrophysical Characterization of a Carbonate Reservoir Pore System in the Offshore Red River Basin of Vietnam 14.

Glover, P., 2012. Petrophysics. United Kingdom: Department of Geology and Petroleum Geology, University of Aberdeen.

Gilbert, M.C. 1982. Geologic setting of the eastern Wichita Mountains with a brief discussion of unresolved problems, in Geology of the eastern Wichita Mountains: Oklahoma Geological Survey Guidebook 21, 1-30.

Gilbert, M.C. 1987. Petrographic and structural evidence from the igneous suite in the

Wichita Mountains bearing on the Cambrian tectonic style of the Southern Oklahoma aul- acogen: Geological Society of America Abstracts with Programs, v. 19, no. 3, 152 p.

Goutorbe, B., Lucazeau, F. and Bonneville, A. 2006. Using neural networks to predict thermal conductivity from geophysical well logs. Geophysical Journal International 166, 115–125.

Gul, S., Aslanoglu, V., Tuzen, M.K. and Senturk, E. 2019. Estimation of Bottom Hole and Formation Temperature by Drilling Fluid Data: A Machine Learning Approach 7.

Gupta, V. 2019. Understanding Feedforward Neural Networks. [online]. Learnopencv [Viewed 17 December 2019] Available from: https://www.learnopencv.com/understanding-feedforward-neural-networks

Ham, W.E., Denison, R.E. and Merritt, C.A. 1964. Basement rocks and structural evolution of southern Oklahoma: Oklahoma Geological Survey Bulletin 95, 302 p.

Ham, W.E. and Wilson, J.L. 1967. Paleozoic epeirogeny and orogeny in the central United States: American Journal of Science, v. 265, 332-407.

Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R.R. 2012. Improving neural networks by preventing co-adaptation of feature detectors. arXiv:1207.0580 [cs].

Huang, K.-Z., Yang, H., King, I. and Lyu, M.R. 2008. Machine Learning: Modeling Data Locally and Globally. Springer-Verlag Berlin Heidelberg. 10.1007/978-3-540-79452-3

Jayalakshmi, T. and Santhakumaran, A. 2011. Statistical Normalization and Back Propagationfor Classification. IJCTE 89–93.

JavaTpoint. 2018. Decision Tree Classification Algorithm. [online]. Javatpoint [Viewed 2 January 2020]. Available from: https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm

Johnson, K. S., Amsde, T.W., Deniso, R. E., Dutton, S. P., Goldstein, A.G., Rascoe, B. Jr., Sutherland P. K. and Thompsan, D. M. 1988, Southern Midcontinenet region, in Sloss, L. L. (ed.), Sedimentary cover – North American craton, U.S.: The Geology of North America, Geological Society of America, Boulder, v. D-2, 307-359.

Jordan, L. and Vosburg, D. L. 1963, Permian salt and associated evaporites in the Anadarko basin of the western Oklahoma-Texas Panhandle region: Oklahoma Geological Survey Bulletin 102, 76 p.

Kansas Geological Survey. 2001.The Hugoton Project, Background. [online]. Kgs.ku.edu. [Viewed 7 October 2019]. Available from: http://www.kgs.ku.edu/Hugoton/background.html

Kleynhans, T., Montanaro, M., Gerace, A. and Kanan, C. 2017. Predicting Top-of-Atmosphere Thermal Radiance Using MERRA-2 Atmospheric Data with Deep Learning. Remote Sensing 9, 1133.

Kok Wai Wong, Chun Che Fung, Yew Soon Ong and Gedeon, T.D. 2005. Reservoir Characterization Using Support Vector Machines, in: International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'06). Presented at the International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'06), IEEE, Vienna, Austria, pp. 354–359.

Komorowski, M., Marshall, D.C., Salciccioli, J.D. and Crutain, Y. 2016. Exploratory Data Analysis, in: Secondary Analysis of Electronic Health Records. Springer International Publishing, Cham, pp. 185–203.

KumarSingh, B., Verma, K., S. and Thoke, A. 2015. Investigations on Impact of Feature Normalization Techniques on Classifier's Performance in Breast Tumor Classification. IJCA 116, 11–15.

Lashin, A. 2005. Reservoir Parameter Estimation Using Well Logging Data and Production History of the Kaldarholt Geothermal Field, S-Iceland 38.

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

Mitchell, T. 2006. The discipline of machine learning.

Natrella M. 2010. Nist/Sematech e-Handbook of Statistical Methods. Nist/Sematech.

Oklahoma Geological Survey., 2008. Earth Sciences and Mineral Resources of Oklahoma: Geologic History of Oklahoma (Educational Publication 8). The University of Oklahoma.

Onalo, D., Adedigba, S., Khan, F., James, L.A. and Butt, S. 2018. Data driven model for sonic well log prediction. Journal of Petroleum Science and Engineering 170, 1022–1037.

Parapuram, G., Mokhtari, M. and Ben Hmida, J. 2018. An Artificially Intelligent Technique to Generate Synthetic Geomechanical Well Logs for the Bakken Formation. Energies 11, 680.

William, J. and Perry, Jr. 1989. Tectonic evolution of the Anadarko basin region, Oklahoma. U.S. Geological Survey bulletin; 1866-A

Prieto, A., Prieto, B., Ortigosa, E.M., Ros, E., Pelayo, F., Ortega, J. and Rojas, I. 2016. Neural networks: an overview of early research, current frameworks and new challenges. Neurocomputing 214, 242–268.

Puskarczyk, E. 2019. Artificial neural networks as a tool for pattern recognition and electrofacies analysis in Polish palaeozoic shale gas formations. Acta Geophys.

Ramcharitar, K. and Hosein, R. 2016. Rock mechanical properties of shallow unconsolidated sandstone. SPE Trinidad Tobago Sect. Energy Resour. Conf.

Rolon, L., Mohaghegh, S.D., Ameri, S., Gaskari, R. and McDaniel, B. 2009. Using artificial neural networks to generate synthetic well logs. Journal of Natural Gas Science and Engineering 1, 118–133.

Schön, D.J. 2015. Basic Well Logging and Formation Evaluation 179

Serra, O., 1984. Fundamentals of well-log interpretation, Developments in petroleum science. Elsevier; Elf Aquitaine, Amsterdam; New York: Pau, 435 p.

Simon, A., Deo, M.S., Venkatesan, S. and Babu, D.R.R. 2015. An Overview of Machine Learning and its Applications 4.

Stanley, T.A., Kirschbaum, D.B., Sobieszczyk, S., Jasinski, M.F., Borak, J.S. and Slaughter, S.L. 2020. Building a landslide hazard indicator with machine learning and land surface models. Environmental Modelling & Software 129, 104692.

Sun, F., Wang, R., Wan, B., Su, Y., Guo, Q., Huang, Y. and Wu, X. 2019. Efficiency of Extreme Gradient Boosting for Imbalanced Land Cover Classification Using an Extended Margin and Disagreement Performance. IJGI 8, 315.

Thongsame, M.W., 2018. Lithological Classification By Deep Learning Algorithm 128.

Tiab, D. and Donaldson, E.C. 2012. Basic Well-Log Interpretation. In Petrophysics, 803-827.

Tomlinson, C.W. and McBee, W., Jr. 1959, Pennsylvanian sediments and orogenies of Ardmore District, Oklahoma, in Petroleum geology of southern Oklahoma, volume 2: American Association of Petroleum Geologists, 3-52.

Ustuner, M., Sanli, F.B. and Abdikan, S. 2016. Balanced vs Imbalanced Training Data: Classifying RapidEye Data with Support Vector Machines. Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci. XLI-B7, 379–384.

Varhaug, M. 2013. Basic Well Log Interpretation.

Wang, G., Cheng, G. and Carr, T. 2013. The application of improved NeuroEvolution of Augmenting Topologies neural network in Marcellus Shale lithofacies prediction. Computers & Geosciences 54, 50–65.

xgboost developers. 2016. Notes on Parameter Tuning — Xgboost 0.83.Dev0 Documentation. [online]. xgboost.readthedocs.io. [Viewed 19 April 2019]. Available from: https://xgboost.readthedocs.io/en/latest/tu torials/param_tuning.html

Xie, Y., Zhu, C., Zhou, W., Li, Z., Liu, X. and Tu, M. 2018. Evaluation of machine learning methods for formation lithology identification: A comparison of tuning processes and model performances. Journal of Petroleum Science and Engineering 160, 182–193.

Zhang, D., Chen, Y. and Meng, J. 2018. Synthetic well logs generation via Recurrent Neural Networks. Petroleum Exploration and Development 45, 629–639.

Zhang, L. and Zhan, C. 2017. Machine Learning in Rock Facies Classification: An Application of XGBoost, in: International Geophysical Conference, Qingdao, China, 17-20 April 2017. Presented at the International Geophysical Conference, Qingdao, China, 17-20 April 2017, Society of Exploration Geophysicists and Chinese Petroleum Society, Qingdao, China, pp. 1371–1374.

Zoveidavianpoor, M., Samsuri, A. and Shadizadeh, S.R. 2013. Prediction of compressional wave velocity by an artificial neural network using some conventional well logs in a carbonate reservoir. J. Geophys. Eng. 10, 045014.