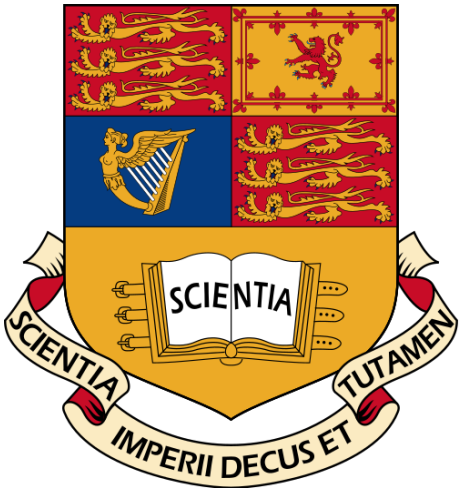


APPLIED COMPUTATIONAL SCIENCE AND ENGINEERING - 2018/2019

IMPERIAL COLLEGE LONDON

MASTER INDEPENDENT RESEARCH PROJECT

Application of Machine Learning to Achieve Robust and Accurate Classification of Well Log Data

---

Author: Runzhi Zhou
Supervisor: Dr. Gerard Gorman, Dr. Navjot Kukreja and Dr. Peter Fitch

Abstract

Machine learning can be used to predict the petrophysical parameters of wells and provide information on facies and fluid properties by well logs data. In this paper, I proposed comprehensive workflows with various machine learning methods, such as logistic regression, SVM, K-means, PCA and t-SNE, that could achieve high accuracies in facies and fluid classification tasks.

From extensive experiments, I observed that: (1) it is fundamental to keep all input features when developing a robust machine learning model; (2) it is easy for both supervised and unsupervised methods to achieve higher prediction accuracies on a small portion of the training samples; (3) supervised learning models are generally more robust and give higher accuracies than unsupervised learning models; (4) I proposed a data augmentation approach by generating more input features with specific powers can improve classification performance remarkably; (5) data splitting according to wells is more robust compared to directly random split the full data set; (6) I proposed a MixedLabel approach for classification, which shows comparable performance.

Acknowledgments

With many thanks to my supervisors, Dr. Gerard Gorman, Dr. Navjot Kukreja, and Dr. Peter Fitch, for their extensive support and guidance throughout this project.

Contents

List of Tables

List of Figures

1 Introduction

1.1 Background and Motivation

In the "Age of Big Data", unravelling unknown data trends and information becomes essential as big data finds its uses and applications in many industries and disciplines, such as in online shopping and bioscience [1, 2]. Big data can be applied in various facets of geoscience industries, for example, in well logging, core fluid analysis, drilling parameters, seismic surveys and so on [3]. Well log data is a form of geoscience data recording data from the log. It is the main tool used by Petro-physicists, geoscientists and engineers to characterise, estimate and predict properties of rocks and fluids in the subsurface [4]. Moreover, as manual well log classification is neither time efficient nor financially considerate, machine learning will be useful in improving prospection efficiencies [5].

Historically, geoscientists have attempted many methods on well log data classification. In 1982, Wolf and Pelissier-Combescure first enabled clustering into electrofacies for geological facies determination and well to well correlation. Wolf and Pelissier-Combescure (1982) defined electrofacies as a set of log responses characterizing sediment [6]. In 1987, J.M. Busch et al. used statistical analysis of wireline log measurements to enable determination of lithology [7]. In 1997, Jong-Se Lim et al. applied hierarchical cluster analysis on well log measurements [8]. In 2001, Trond Mathisen et.al successfully applied non-parametric regression techniques to characterize electrofacies with permeability prediction [9]. In 2009, Y. Zee Ma analysed the popular artificial neural network (ANN) for lithology or facies clustering [10]. Recently, Yunxin et al. compare different supervised learning applications on well log [11]. Researchers are also exploring unsupervised learning methods on well log data [12,13].

Geoscientists have proved that both workflow and methods could impact subsurface analysis results for lithology, fluid and facies clustering, therefore exploring different workflows and machine learning methods on well log data becomes this scholar's interest [4]. Many different and complex supervised and unsupervised classification techniques can be applied on well logs, however, considering that this scholar only have 6 independent features and a limited amount of training data, this essay uses simpler classification and clustering techniques such as K-means, logistic regression, SVM, PCA, and t-SNE. Supervised learning could overfit due to limited number of labelled data set. Choosing a suitable data augmentation method can alleviate this issue [2].

Noticeably, different geologists might come up with different labelling schemes on well log data. Moreover, difference in input feature types for well log data sets makes hyper-parameter impossible to be applied from one data set to another. As we cannot simply compare accuracy across data sets, this paper aims to establish standard procedures with different machine learning methods on all well log data sets to achieve robust and accurate classification.

1.2 Contributions

This essay starts with revision on basic machine learning methods and technical platforms in chapter 2. This essay then introduces the data set and data splitting methods employed in chapter 3. And chapter 4 explores supervised learning methods. In this chapter, logistic regression is used to select input features, following by comparison on different data splitting methods. In chapter 4, this

scholar proposes a new augmentation method and introduces threshold methods on logistic regression. Chapter 5 explores unsupervised learning methods by using PCA [14] and t-SNE [15] as qualitative evaluation methods to generate two-dimensional representations, and K-means clustering as quantitative evaluation method.

2 Basic Methods

Supervised learning and unsupervised learning are two categories of machine learning. If the data are given with labels, then the learning is supervised. In contrast, unlabelled data can only do unsupervised learning [16]. In this work, I only have 6 independent features and a limited amount of training data, so we decided to rely on simpler classification techniques, for example, K-means, PCA [14], t-SNE [15] Logistic Regression, and SVM [17].

2.1 Dimensionality Reduction

PCA and t-SNE are used to represent high-dimensional features in lower dimensions. I will make 2-dimension plots in this paper to visually assess similarities and differences between samples and perform clustering.

2.1.1 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) [14] is a linear feature extraction technique, which performs a linear mapping of the data to a lower-dimensional space, that maximises the variance of the data in low-dimensional representation by calculating the eigenvectors from the covariance matrix.

2.1.2 t-Distributed Stochastic Neighbour Embedding (t-SNE)

In contrast to PCA, *t-Distributed Stochastic Neighbor Embedding (t-SNE)* [15]is a non-linear dimension reduction technique. t-SNE can analyse the multi-dimensional data to a lower dimensional space trying to find patterns in the data by identifying observed clusters based on the similarity of data points with multiple features. t-SNE is mainly both a data exploration and visualization technique, which can become more robust with respect to the presence of outliers [18].

2.2 Supervised Learning

2.2.1 Logistic Regression

Logistic regression is a supervised machine learning method. The logistic function, also called the sigmoid function, has an S-shaped curve that can generate any real-valued number into a number between 0 and 1, but never exactly at those limits [19].

Since a logistic function is monotonically increasing, the absolute values of weights determine the importance of features. We can therefore find the importance of each inputs features by comparing the absolute values of weights.

2.2.2 Support-Vector Machines (SVM)

Support-Vector Machines (SVM) [17] are supervised learning models that could build optimal separating boundaries between data sets by solving a constrained quadratic optimization problem. We tried our purposed data augmentation methods on both Logistic Regression and SVM to get a better conclusion.

2.3 Unsupervised Learning

2.3.1 *K*-means Clustering

Data Clustering is used for grouping unlabelled samples (observations, data items, or feature vectors). K-means clustering is an unsupervised machine learning method [20]. The basic idea is first to get K centers of the K clusters (minimise the inertia) by k-means by using training data and then to use those clusters center to predict the classes of the testing data.

2.4 Code Metadata

We have already open sourced our code on Github (code).

2.4.1 Technical Platform

Jupyter notebooks Python 3.7.3
Anaconda version: Anaconda 4.6.14
Operating System: Windows 10

2.4.2 Programming Language

Python 3.7.3

2.4.3 Dependencies

See Table 1

Library	**Packages**	**Usages**	**Links**
sklearn 0.20.3	preprocessing	Data normalization	https://scikit-learn.org
sklearn 0.20.3	cluster	K-means	https://scikit-learn.org
sklearn 0.20.3	SVC	SVM	https://scikit-learn.org
sklearn 0.20.3	t-SNE	t-SNE	https://scikit-learn.org
pandas 0.24.2	Dataframe	Change data form	http://pandas.pydata.org
pandas 0.24.2	Series	Select data	http://pandas.pydata.org
numpy 1.16.2	ndarray	Select and compute data	https://www.numpy.org
pytorch 1.1.0	nn;optim	Used for neural networks	https://pytorch.org
matplotlib 3.0.3	pyplot	Ploting	https://matplotlib.org
openpyxl	load_workbook	Read excel files	http://openpyxl.readthedocs.io
livelossplot	PlotLosses	Neural networks loss	https://github.com/stared/livelossplot

Table 1: The library with usages and links

3 Data Processing

The data used throughout this project comprised of a suite of logging data obtained for the Wytch Farm oilfield. Wytch Farm is the largest onshore oilfield in Western Europe and is situated on the south coast of the UK, beneath Poole Harbour and the Isle of Purbeck. There are five wells with log data from our dataset. (well 1: 1D-02) (well 2: 1F-11) (well 3: 1K-01) (well 4: 1X-02) (well 5: 98_6-8) (see Figure 1)

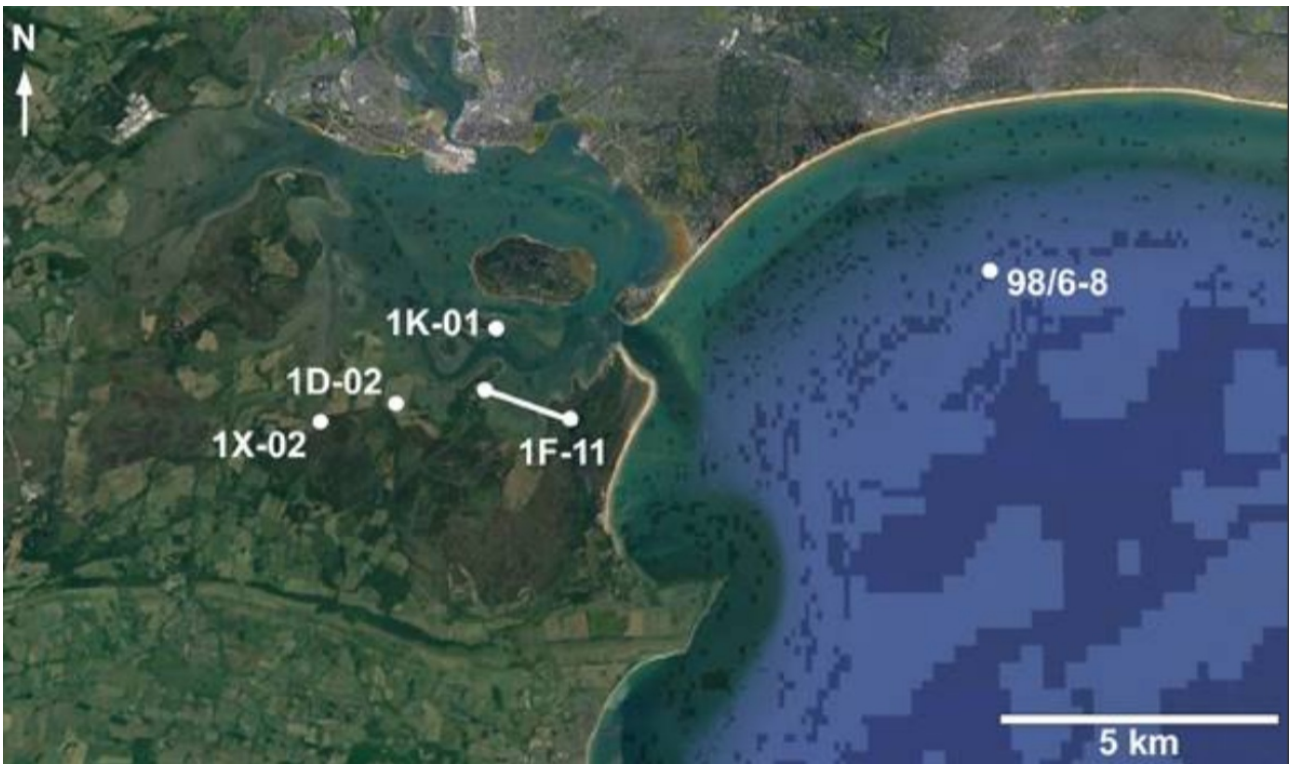


Figure 1: The locations of all five wells in Wytch Farm (Reference to Fitch, 2019).

3.1 Data

3.1.1 Data Set

Nine different input well log features and "labels" are provided for five wells (1D-02, 1F-11, 1K-01, 1X-02 and 98_6-8). There are 9 independent features across all data but only 6 of them (*Depth, DT, GR, LLD, NPHI, RHOB*) are common features. We list the meaning of each input feature in the Table 2.

Names	Units	Meanings
Depth	m	Depth down the borehole;
CALI	inches	How wide the borehole is;
DT	sec/ft	P-wave Transit Time (or sonic log);
GR	API	Gamma Ray;
LLD	ohm.m	Deep Resistivity;
LLS	ohm.m	Shallow Resistivity;
MSFL	ohm.m	Micro Resistivity;
NPHI	pu (porosity units)	Neutron Porosity;
RHOB	g/cm^3	Bulk Density;

Table 2: Nine input features with their meanings.

3.1.2 Data Normalization

The raw input features are of different magnitude. For example, in our data set, Gamma Ray may typically varies between 50 and 200 g API, whereas NPHI might be expected to fall in the range of 0 – 0.5 pu. In order to avoid any feature to dominating models' decisions, I exploited a standard Gaussian normalization method for all the input data. The standard score Y of a sample x is calculated as:

$$Y = (x - u)/s$$

where u is the mean of the training samples and s is the standard deviation of the training samples Method.

3.2 Labels

There are two original label schemes for all the data in our data set, *i.t.,* 'Log_facies' and 'Fluids'. I proposed a new label scheme (*'MixedLabel'*) by mixing 'Log_facies' and 'Fluids'. Details of scheme meanings of each label scheme will be explained in Table 3 and Table 4. We visualize data with labels in Figure 2. We develop the 'MixedLabel' scheme for two reason. Firstly, we try to get more detailed label for one classification. Secondly, we want to check whether models trained with mixed label can get higher accuracy or not.

3.3 Training and Testing

There are five labeled wells with total 4695 data set. We use three different training and testing methods for those data.

3.3.1 Single Well with Cross-validation Method

The data from a single well in our data set are randomly split with 90 percent used as the training set and 10 percent used as the testing set. The accuracy is obtained from averaging results from 10 independent trials.

Scheme Names	Meanings
Log_Facies (Scheme)	These exist for the five wells predicted by the geologist so that they can predict facies in the un-cored well based on the log signature; 3 categories; (1: channel sandstone) (2: floodplain mudstone) (3: lacustrine and paleosol mud)
Fluid (Scheme)	These exist for the five wells predicted by the geologist; 2 categories; (1: in the hydrocarbon-bearing zone) The mixture of hydrocarbon and formation water; (2: in the water-bearing zone) Only formation water;
MixedLabel (Scheme)	These exist for the five wells Mixed by the prediction of Log_Facies and Fluid by the geologist; 5 categories; (1: - Log_Facies 1 & Fluid 1) (2: - Log_Facies 1 & Fluid 2) (3: - Log_Facies 2 & Fluid 1) (4: - Log_Facies 2 & Fluid 2) (5: - Log_Facies 3 & Fluid 1) Noticeably, there is always 5 categories in this MixedLable scheme since there is no Log_Facies 3 & Fluid 2 categories throw all of our data;
WellNumber	These exist for the 5 wells from our dataset; (1: 1D-02) (2: 1F-11) (3: 1K-01) (4: 1X-02) (5: 98_6-8)

Table 3: Original schemes and mixed label schemes.

Facies Names	Log Signature	Uncertainty
1-channel sandstone	RHOB to the left of NPHI; Highly variable separation of RHOB/NPHI; Higher porosity and permeability; Wide range of porosity and permeability	Not able to use gamma-ray to distinguish between grain size trends in the sandstones because of the high K content so unable to separate finer channel sands from rest of channel fill; Resolution of RHOB/NPHI logs not high enough to be able to see cemented lags;
2–floodplain mudstone	RHOB to the right of NPHI; Smaller separation of RHOB/NPHI; Low porosity and permeability; Wide range of porosity and permeability;	Core observations shows that this mudstone is often interbedded with smaller sand lenses. The resolution of the log doesn't allow them to be separated out; Also groupings were made at a resolution of 1m or more and so some very small sand bodies have had to be correlated in this log facies;
3–lacustrine and paleosol mud	RHOB to the right of NPHI; Large RHOB/NPHI separation; Very low porosity and permeability; Smaller range of porosity and permeability;	Grouped together lacustrine and paleosol/top of floodplain core facies based on similar log signatures as I recognized uncertainty in the core interpretation as they were very hard to tell apart in the core;

Table 4: Details of Log_Facies scheme

3.3.2 Multi Wells with Cross-validation Method

The data from all the wells in our data set are randomly split with 90 percent used as the training set and 10 percent used as the testing set. The accuracy is obtained from averaging results from 10 independent trials. We only used the 6 common inputs features for multi well testing.

3.3.3 Multi Wells with Cross-wells Method

Select four wells in our data set as training data and use the remaining well in data set 1 as testing data. Run 5 independent trials, each with a different testing well, to make sure all the five wells have been used as a testing data once. Calculate the average accuracy. I only used the 6 common input features for multi well testing.

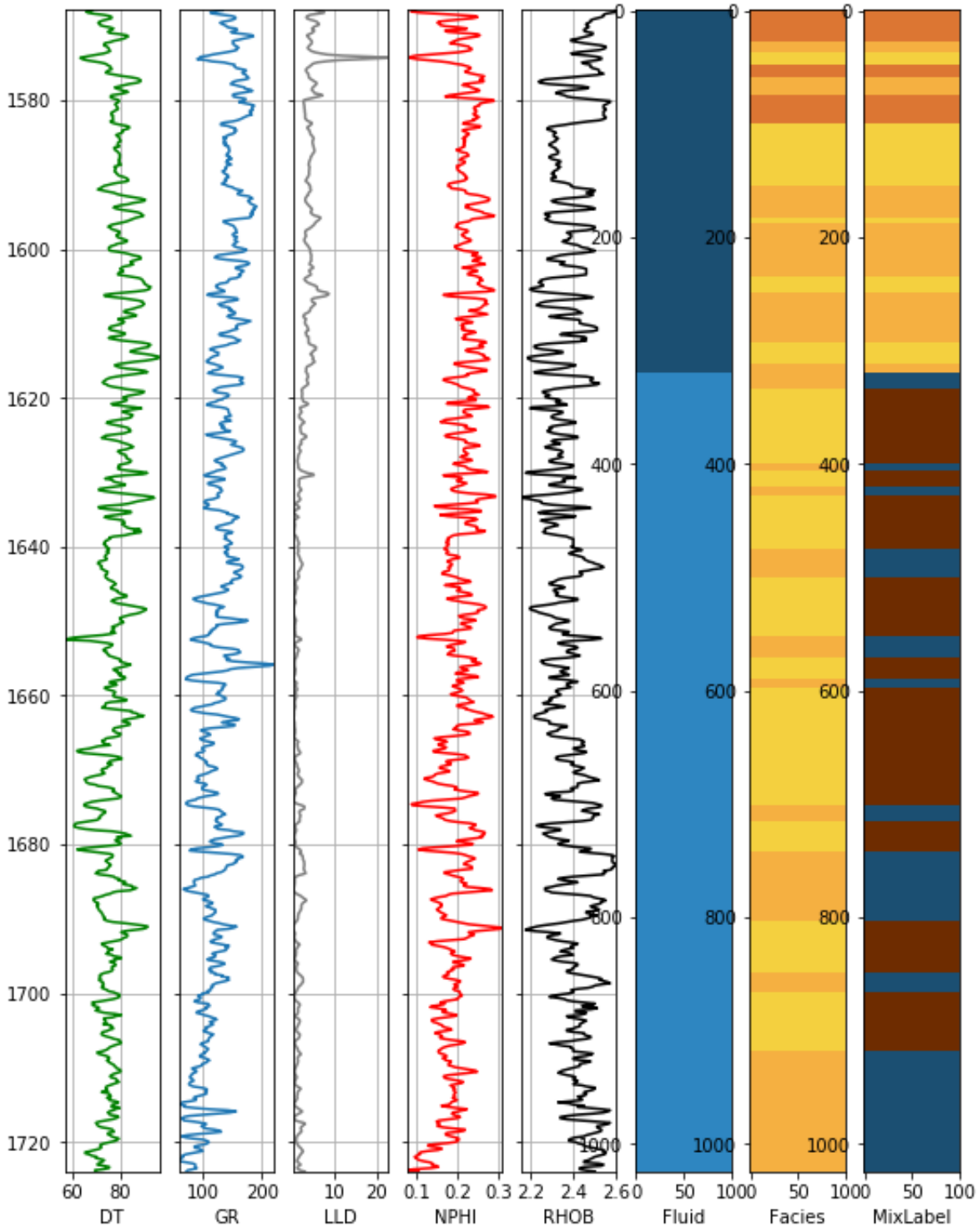


Figure 2: Example of well logs data from one well(1K-01) with three different label schemes (*Implementation Code*).

4 Supervised learning approach and evaluation

In this chapter, we explored the well-known supervised learning approaches, i.e., logistic regression and SVM on well log classification. We firstly analysed the influence of each feature; Secondly, we discussed the best training and testing method; Thirdly, we proposed an augmentation method to improve the accuracy; Finally, we tried a method based on logistic regression with threshold prediction rate to achieve higher accuracy on small group of data set. See our code from the link.

4.1 Inputs Features Selection

We designed two series of experiments based on logistic regression: 1) we analyzed the influence of each feature; 2) we trained with all features. See our code from the link.

From the Table 5, for the Fluid, the absence of LLD feature lower the accuracy of testing; for the Log_Facies, the absence of NPHI and RHOB lowering the accuracy of testing; for the MixedLabel, the absence of LLD, NPHI and RHOB all lowering the accuracy of testing. Moreover, the accuracy of the 'MixedLabel' is always approximately the product of the accuracy of 'Fluid' and the accuracy of 'Log_Facies'. Therefore, the decrease of the accuracy of 'Fluid' or accuracy of 'Log_Facies' may result in a decrease of the accuracy of the 'MixedLabel'. This may conclude that mixing the labels is not helpful for supervised learning.

Moreover, since logistic function is monotonically increasing, the absolute values of weights determine the importance of features. We can therefore find the importance of each input features by comparing the L1-norm of each feature(sum of absolute values of three weights of each feature). We observe that the L1-norm of 'Depth' and 'RHOB' are large from both table 6 and 7. However, the L1-norm of 'Depth' feature become very small in table 8, while the L1-norm of NPHI and RHOB features keep high.

As mentioned in section 3.2, 'NPHI' and 'RHOB' are two of the most important features for 'Log_Facies' classification, which is same as shown in table 8 for multi wells. 'Depth' seems to play an important role for single well model. Therefore, we ran new experiment to explore whether 'Depth' is a significant input feature. We choose those two wells because the depth of well 1F-11 is in a different range compare to that of well 1K-01.

The table 9 shows that depth will decrease the accuracy of all 'Fuild', 'Log_Facies' and 'MixedLabel'. It means that the depth may have negative influence for the accuracy test for classification predictions when the training set is small. However, when the training set is relatively large, as shown in table 5, depth feature does not have any negative effect.

4.2 Training and Testing

In this section, we ran three types of experiments (as mentioned in section 3.3): 1) we use the data from single wells with cross-validation method; 2) we use the data from multi wells with cross-validation method; 3) we use the data from multi wells with cross wells method.

	Fluid	Log_Facies	MixedLabel
Accuracy with all features	0.95	0.82	0.78
Accuracy without 'DEPTH'	0.94	0.81	0.78
Accuracy without DT	0.94	0.81	0.79
Accuracy without GR	0.94	0.81	0.77
Accuracy without LLD	0.82	0.81	0.67
Accuracy without NPHI	0.95	0.75	0.73
Accuracy without RHOB	0.95	0.68	0.68

Table 5: Multi wells accuracy of logistic regression with cross validation method (*Implementation Code*).

	DEPTH	CALI	DT	GR	LLD	LLS	MSFL	NPHI	RHOB
Class-1 weights	0.0654	-0.2049	-0.992	0.2787	1.0139	-1.5144	1.1572	-1.2475	-3.5645
Class-2 weights	1.1372	-0.0303	0.6867	0.0546	0.9445	-0.1239	-0.6896	0.4112	1.6887
Class-3 weights	-3.2547	1.0967	0.0994	0.0892	-1.3322	0.8375	0.0253	0.2457	1.0398
L1-norm	4.4573	1.3319	1.7781	0.4225	3.2906	2.4758	1.8721	1.9044	6.2930

Table 6: The weights and its L1-norm of logistic regression for well 1K-01 with 9 inputs features. The feature weights with the largest two L1-norm are marked as red (*Implementation Code*).

	DEPTH	DT	GR	LLD	NPHI	RHOB
Class-1 weights	-1.0054	0.2752	-0.3944	0.9254	-1.2766	-3.7828
Class-2 weights	0.7984	-0.4165	0.1919	-0.7669	0.1980	1.8236
Class-3 weights	-0.4231	1.203	0.5706	0.0095	0.4360	1.1768
L1-norm	2.2269	1.8947	1.1569	1.7018	1.9106	6.7832

Table 7: The weights and its L1-norm of logistic regression for well 1K-01 with 6 common inputs features throughout dataset 1. The feature weights with the largest two L1-norm are marked as red (*Implementation Code*).

	DEPTH	DT	GR	LLD	NPHI	RHOB
Class-1 weights	0.3626	-0.1178	-0.0185	-0.1905	-2.6303	-4.0262
Class-2 weights	-0.0963	-0.0890	-0.2362	-0.0651	1.2637	1.9491
Class-3 weights	-0.9516	0.1216	1.1952	0.6124	0.4884	1.4670
L1-norm	1.4105	0.3284	1.4499	0.8680	4.3824	7.4423

Table 8: The weights and its L1-norm of logistic regression for all the wells in our dataset with 6 common inputs features throughout our data set. The feature weights with the largest two L1-norm are marked as red (*Implementation Code*).

	Fluid	Log_Facies	MixedLable
Accuracy with All Features	0.31	0.06	0.10
Accuracy without 'DEPTH'	0.68	0.55	0.46
Accuracy without 'DT'	0.31	0.33	0.10
Accuracy without 'GR'	0.31	0.06	0.10
Accuracy without 'LLD'	0.31	0.06	0.10
Accuracy without 'NPHI'	0.31	0.07	0.10
Accuracy without 'RHOB'	0.31	0.06	0.07

Table 9: Multi wells accuracy of logistic regression with cross validation method (*Implementation Code*).

From table 10, most of the accuracy of single wells cross validation method is higher than the Accuracy of multi wells cross validation method. The Accuracy of multi wells cross wells method is the lowest.

The high accuracy of single well cross validation may due to the fact that data was collected from the same well as the training data of the model, so it is very similar to what the model has trained before. Moreover, the relatively higher accuracy of multi wells cross validation method than that of multi wells cross wells method should have been caused by the same reason.

Methods	Well/Wells	Fluid	Log_Facies	MixedLabel
Single well with cross validation method	1D-02	0.99	0.87	0.88
Single well with cross validation method	1F-11	0.98	0.90	0.88
Single well with cross validation method	1K-01	0.98	0.77	0.80
Single well with cross validation method	1X-02	0.99	0.86	0.84
Single well with cross validation method	98_6-8	0.99	0.82	0.79
Multi wells with cross validation method	all 5 wells	0.96	0.81	0.79
Multi wells with cross wells method	all 5 wells	0.89	0.80	0.71

Table 10: The accuracy of different data selection run throw logistic regression with different training and testing methods (*Implementation Code*).

4.3 Our Proposed Data Augmentation Method

From the study of the data distribution shown in Figure 3, we noticed that the mode of the range of each feature is helpful to divide different class. Therefore, I propose to add more features by adding powers of the raw data to increase separations of class distributions. We performed the same data normalization approach as shown in the previous section after the data augmentation step. See the code from the link.

We plan to generate new features from the raw features to create new data set with more features. As shown in figure 4, We generate five new features in a set each time (discard 'Depth' as it is uniformly distributed) and add those features set to the original data set (6 features) separately to get several new data sets with 11 features each. I generated new features by setting the power from the range between -20 and 20 with the step of 0.5, and ran experiments separately based on logistic regression and SVM.

From Figure 5, we can see that the performance consistently improved with both logistic regression and SVM after using our proposed data augmentation method.

4.4 Setting Threshold Prediction Rate

We use a SoftMax layer for all the supervised learning method. However, SoftMax give a prediction based on the highest predicted rate only. I would like to set a threshold percentage for the original SoftMax layer that only those whose predicted class score is higher than the threshold are classified. See the code from the link. Since there might be some predicted rate lower than the threshold percentage, we introduce two definitions: Threshold predicted rate(TPR) and Threshold predicted accuracy(TPA).

(1) Threshold Predicted Rate (TPR):

TPR: the percentage of tested data that have a prediction of classes

TPR = tested data that have a prediction of classes/ total tested data

(2) Threshold Predicted Accuracy (TPA):

TPA: The accuracy of all the tested data that have a prediction of classes

TPA = correct predicted tested data that have a prediction of classes/ total tested data that have a prediction of classes.

Table 11 shows that when the threshold of the SoftMax layer increases, the threshold predicted rate will decrease and the threshold predicted accuracy will increase. If we set a relatively high threshold, we can always achieve high classification accuracy in terms of a small portion of data set.

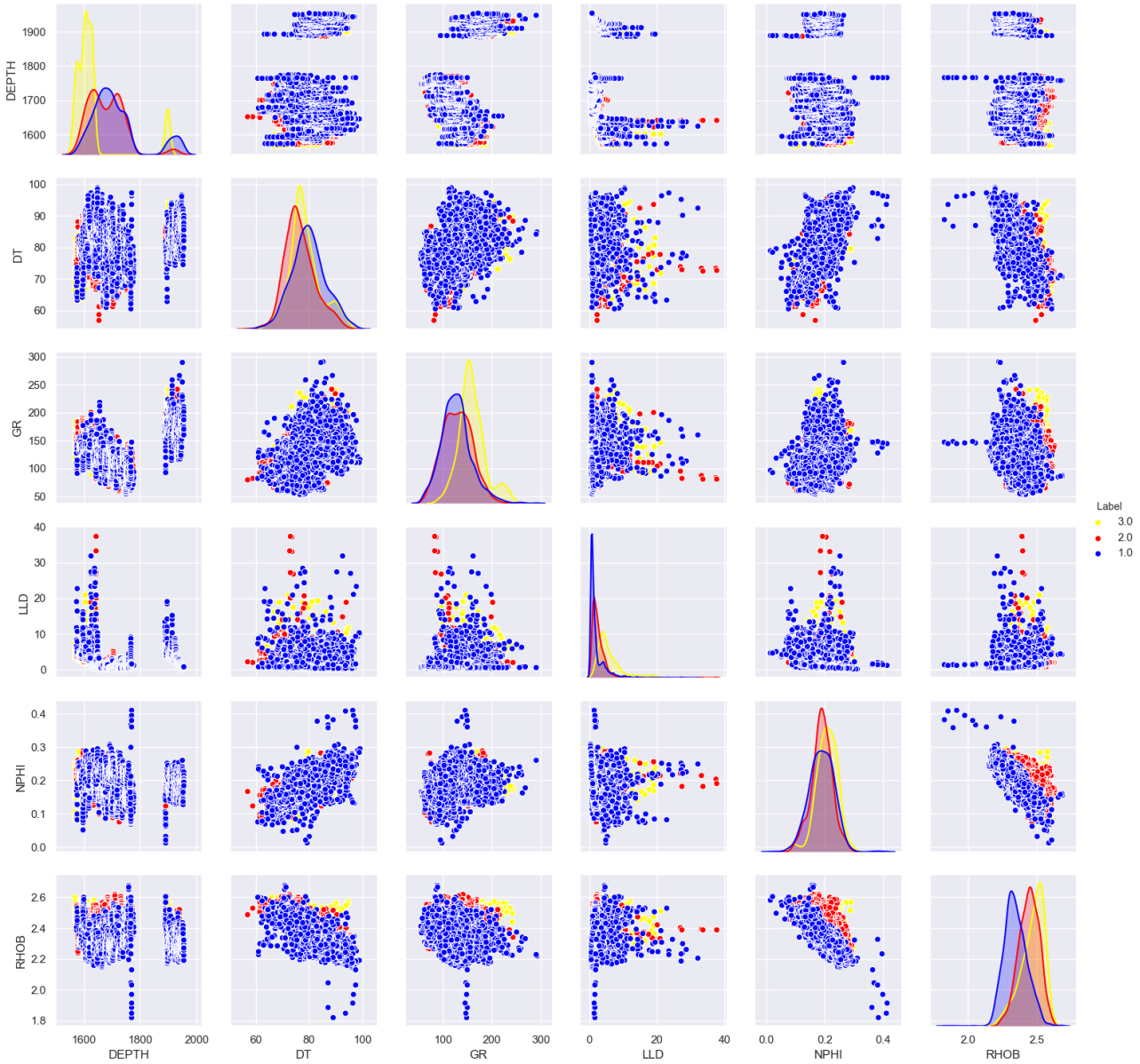


Figure 3: The correlations graphs between six common inputs features from data set 1 one by one (Log_Facies1:blue; Log_Facies2:red; Log_Facies3:yellow) (*Implementation Code*).

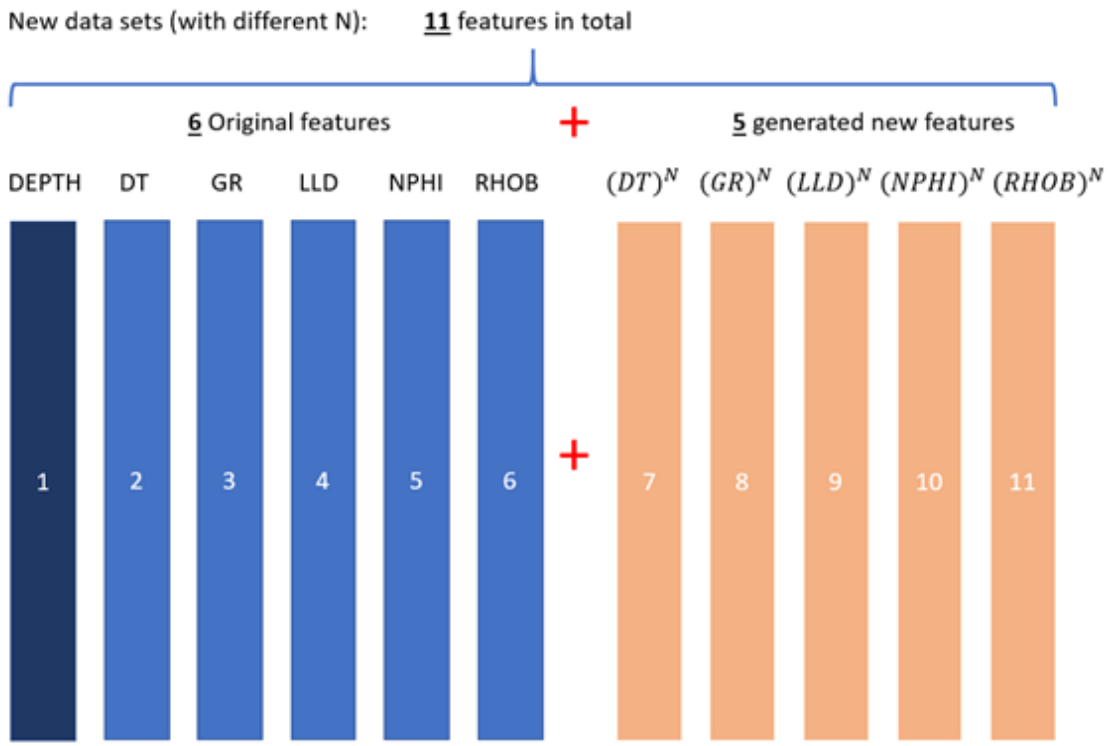
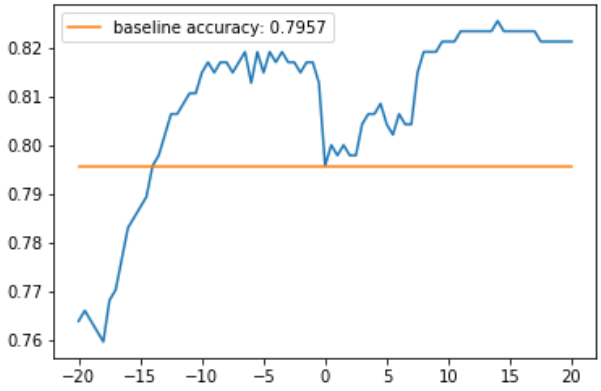
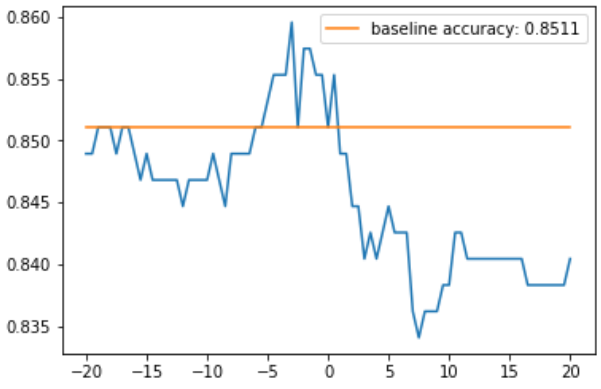


Figure 4: Workflow of our data augmentation method (creating new data sets with 11 features).



(a) Logistic Regression



(b) SVM

Figure 5: Accuracy changes test by logistic regression (left) and SVM (right). X-axis represents the values of power. Y-axis represents the corresponding testing accuracy. The orange line represents the baseline result with raw features. *Implementation Code*

Threshold / Accuracy	Softmax	50%	60%	70%	80%	90%	95%
Fluid TPR	1	1	0.97	0.94	0.92	0.87	0.80
Fluid TPA	0.95	0.95	0.97	0.98	0.98	0.99	0.99
Log_Facies TPR	1	0.98	0.86	0.71	0.52	0.25	0.09
Log_Facies TPA	0.81	0.82	0.86	0.91	0.95	0.98	0.98
Mixedlabel TPR	1	0.87	0.70	0.53	0.32	0.05	0.01
Mixedlabel TPA	0.78	0.83	0.88	0.91	0.95	0.95	0.97

Table 11: Threshold predicted rate(TPR) and Threshold predicted accuracy(TPA) for three different labels schemes (*Implementation Code*).

5 Unsupervised Learning Approaches and Evaluation

In this chapter, we firstly visualized the well log data using PCA and T-SNE and then we used K-means clustering to quantitatively evaluate the classification performance. See our code from the link.

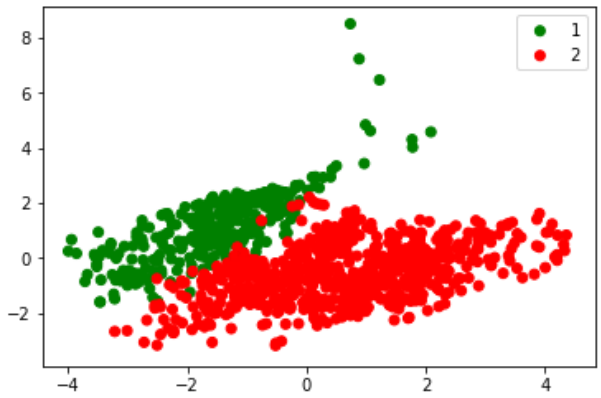
5.1 Qualitative Evaluation by PCA and t-SNE

This part will show the 2D results for selecting facies (Log_Facies), fluids (Fluid), facies with fluids (MixLabel) and different wells separately.

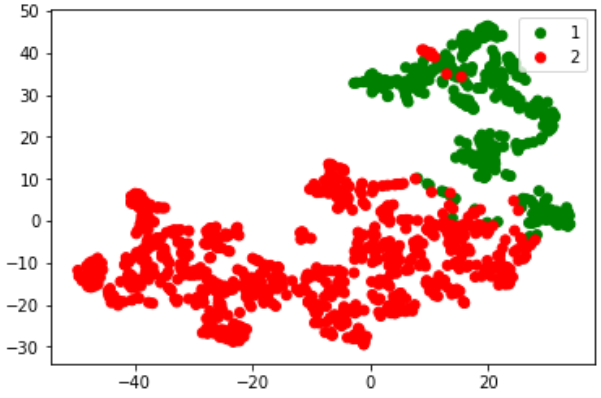
The Figure 6 shows the visualization of high-dimensional data from a single well with PCA and t-SNE methods for Fluids, Log_Facies and MixedLabel labels. While data labeled as 'Fluid' can be clearly separated into two groups by both methods (see Figure 6 (a)(b)), it is hard to differentiate well log data with more than two classes by using PCA and t-SNE (see Figure 6 (c)(d)(e)(f)).

The Figure 7 visualizes high-dimensional data from multi wells with PCA and t-SNE method for Log_Facies label and all wells. Figures 7 (a)(b) shows that 'Log_facies' classes are mixed together and it is hard to distinguish any of the classes. Neither PCA nor TSNE can divide the classes. It shows unsupervised learning is not suitable for this task, especially for the label of 'Log_Facies'.

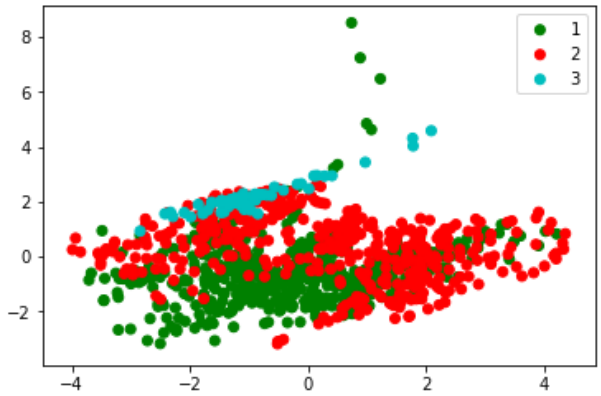
Figures 7 (c)(d) shows that one well is divided into another cluster. This suggests that unsupervised learning may be useful to divide wells. Figure 7 (e)(f) shows that we cannot separate wells if we do not include the Depth feature. As a result, 'Depth' is a dominate feature for different wells while lowering dimension. All other features cannot easily achieve data separations. All the figures in this section further show that unsupervised learning may not perform well for classification.



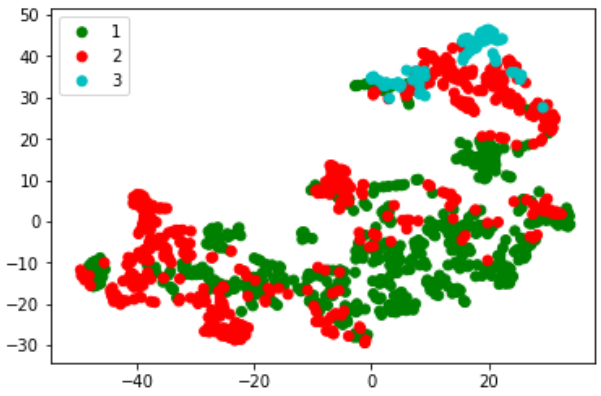
(a) Fluid 1: green dot; Fluid 2: red dot. (PCA)



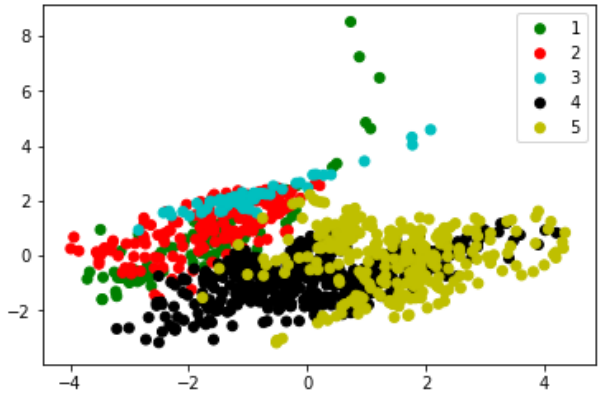
(b) Fluid 1: green dot; Fluid 2: red dot. (t-SNE)



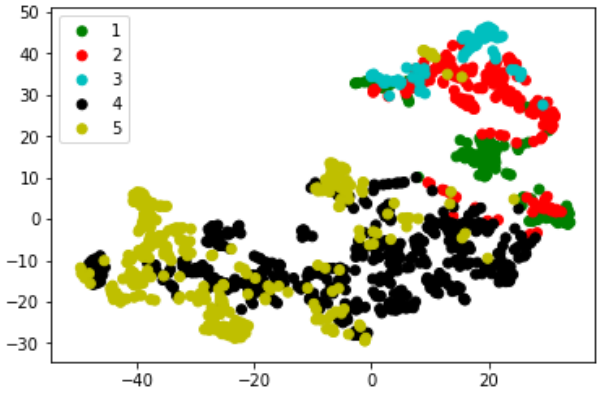
(c) Log_Facies 1: green dot; Log_Facies 2: red dot; Log_Facies 3: light blue dot. (PCA)



(d) Log_Facies 1: green dot; Log_Facies 2: red dot; Log_Facies 3: light blue dot. (t-SNE)

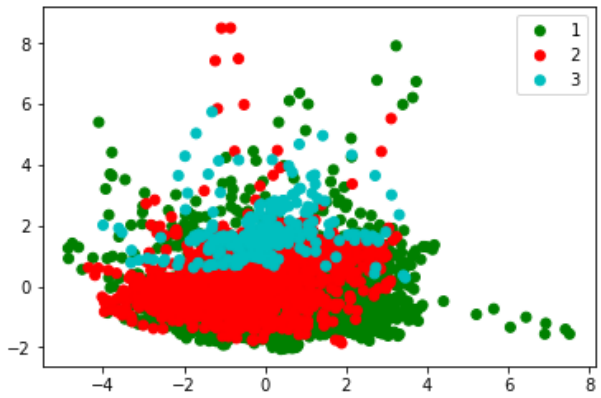


(e) MixedLabel 1: green dot; MixedLabel 2: red dot; MixedLabel 3: light blue dot; MixedLabel 4: black dot; MixedLabel 5: yellow dot. (PCA)

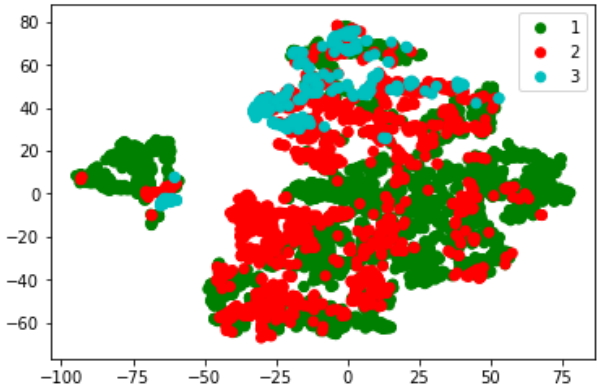


(f) MixedLabel 1: green dot; MixedLabel 2: red dot; MixedLabel 3: light blue dot; MixedLabel 4: black dot; MixedLabel 5: yellow dot. (t-SNE)

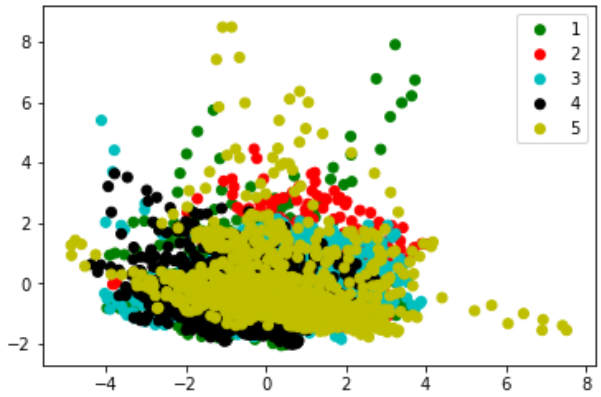
Figure 6: Visualization with PCA (left) and T-SNE (right) based on well log data from a single well(1K-01). 6 common features (e.g., Depth, DT, GR, LLD, NPHI, RHOB) are used as input (*Implementation Code*).



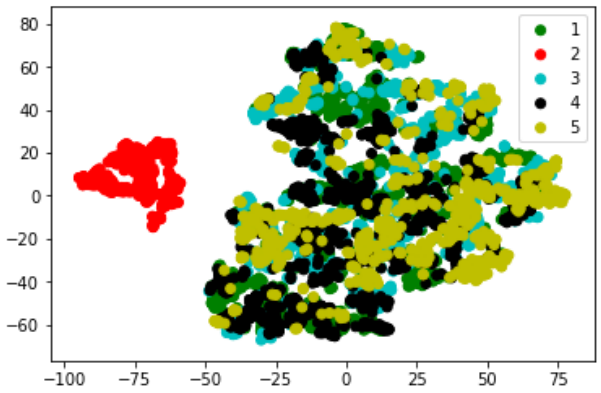
(a) Log_Facies 1: green dot; Log_Facies 2: red dot; Log_Facies 3: light blue dot. (PCA)



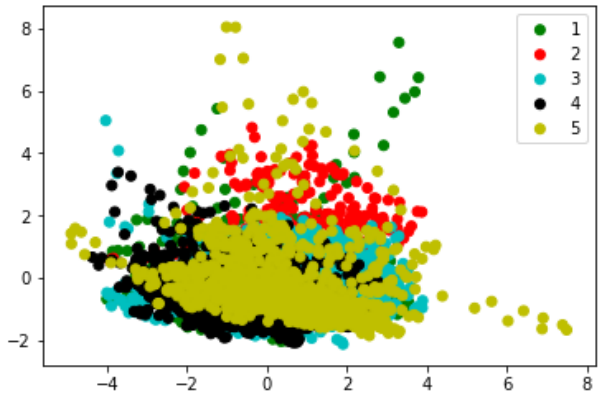
(b) Log_Facies 1: green dot; Log_Facies 2: red dot; Log_Facies 3: light blue dot. (t-SNE)



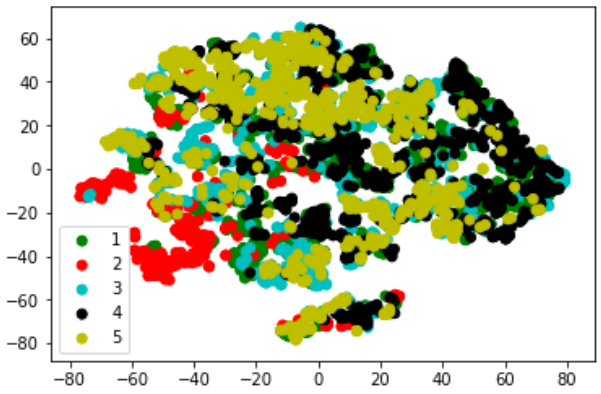
(c) Well 1: green dot; Well 2: red dot; Well 3: light blue dot; Well 4: black dot; Well 5: yellow dot.(PCA)



(d) Well 1:green dot; Well 2:red dot; Well 3:light blue dot; Well 4:black dot; Well 5:yellow dot.(t-SNE)



(e) Well 1: green dot; Well 2: red dot; Well 3: light blue dot; Well 4: black dot; Well 5: yellow dot.(PCA)



(f) Well 1:green dot; Well 2:red dot; Well 3:light blue dot; Well 4:black dot; Well 5:yellow dot.(t-SNE)

Figure 7: Visualization with PCA (left) and T-SNE (right) based on well log data from multiple wells. 6 common features (i.e., Depth, DT, GR, LLD, NPHI, RHOB) are used as input for figures (a, b, c, d). 5 common features (i.e., DT, GR, LLD, NPHI, RHOB) are used as input for figures (e, f) (*Implementation Code*).

K	2	3	4	5	6	7	8	9	10
Score	18772	14867	12373	10749	9326	8785	8232	7768	7316

Table 12: The corresponding score (of the Sum of squared distances of samples to their closest cluster center) for each K (*Implementation Code*).

5.2 Quantitative Evaluation with K-means Clustering

In this part, we designed experiments to find a sufficient number of clusters which allows us to achieve high classification accuracy on well log dataset with K-means clustering approach.

From the Table 12 and the Figure 8, the sum of squared distances of samples to their closest cluster centre decreases dramatically until the number of clusters reaches to 6. Therefore, $k = 6$ should be a sufficient number of clusters, and we will use this value to do quantitatively unsupervised learning experiments.

Table 13 shows that the clustering method can always get some cluster with high accuracy. However, some of the clusters have low accuracy.

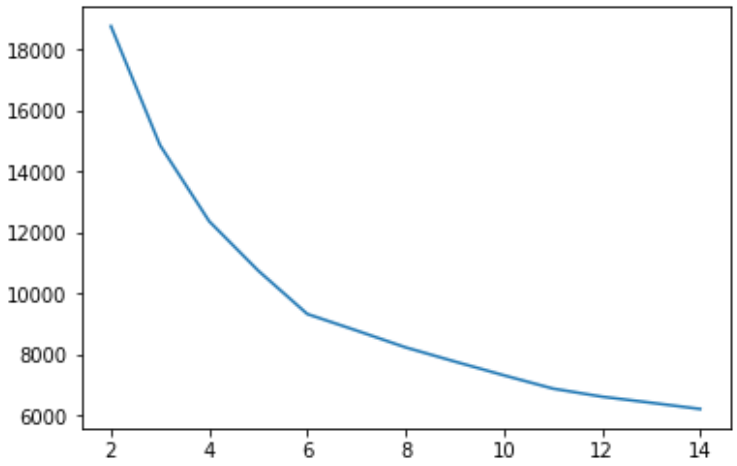


Figure 8: Sum of squared distances of samples to their closest cluster(Y-axis) center versus K value(X-axis) (*Implementation Code*).

Schemes	Classes	K	Cluster Labels	Labelled Clusters Accuracy	Total Accuracy
Fluid	2	6	Cluster1 (Fluid-1) Cluster2 (Fluid-1) Cluster3 (Fluid-1) Cluster4 (Fluid-1) Cluster5 (Fluid-2) Cluster6 (Fluid-1)	Cluster1 (0.93) Cluster2 (0.82) Cluster3 (0.63) Cluster4 (0.73) Cluster5 (0.99) Cluster6 (1.00)	0.85
Log_Facies	3	6	Cluster1 (Log_Facies-1) Cluster2 (Log_Facies-1) Cluster3 (Log_Facies-1) Cluster4 (Log_Facies-1) Cluster5 (Log_Facies-2) Cluster6 (Log_Facies-1)	Cluster1 (0.81) Cluster2 (0.47) Cluster3 (0.78) Cluster4 (0.78) Cluster5 (0.51) Cluster6 (0.67)	0.67
MixedLabel	5	6	Cluster1 (MixedLabel-1) Cluster2 (MixedLabel-4) Cluster3 (MixedLabel-1) Cluster4 (MixedLabel-4) Cluster5 (MixedLabel-4) Cluster6 (MixedLabel-2)	Cluster1 (0.67) Cluster2 (0.77) Cluster3 (0.54) Cluster4 (0.58) Cluster5 (0.46) Cluster6 (0.40)	0.58

Table 13: Unsupervised accuracy by sufficient K=6. Cluster labels represent the cluster index and the majority of true label in the corresponding clusters (*Implementation Code*).

6 Discussion and Conclusions

In this paper, we firstly demonstrate the importance of 6 independent features, and we stress that each feature dimension plays a vital role in differentiating classes, *i.e..,* Log_Facies, Fluid. Specifically, *ROHGB* and *NPHI* are closely related to the supervised classification performance. *Depth* will become a dominant feature for classification in the case where only a small amount of training samples are provided. However, when we use a relatively larger training data set (*e.g.* 5 wells), the influence from *Depth* decreases remarkably. Therefore, if we have sufficient training samples, it is worth to keep as many inputs features as possible to build a robust supervised model. We also show that mixing facies with fluids as new label is not useful for both supervised and unsupervised classification.

We then show that both the supervised and unsupervised models can achieve high prediction accuracy on a small portion of data set. There might be some samples with obvious features for classifications among all the data. Noticeably, the supervised learning methods got a higher accuracy (with total accuracy and threshold accuracy) than unsupervised learning methods. Furthermore, since it is inevitably to use labelled data even if for unsupervised learning approaches, these methods do not show clear advantages compared with supervised learning methods.

In this paper, we proposed a new data augmentation method which shows advantageous in improving classification performance with both logistic regression and SVM models. It worth noting that we have only tried to add one set of input features for these tests. In the future study, we would try other feature generation approaches.

Even though our models provide a high classification accuracy on each label, it is not yet sufficient to convince its robustness and reliability regarding larger dataset as we do not have enough training dataset, which is a key factor to build a robust machine learning models. However, some clear workflows have demonstrated their power in handling this task through this report. Noticeably, in general, we should make prediction on new wells, which are unseen during the training process. In this way, the trained machine learning models can be considered to have the generalizing capacity. Therefore, dataset split based on cross wells should be the most meaningful splitting approach.

In conclusion, all the dimensions of input features play a vital role in well log data classification. While unsupervised learning methods have the potential to train the models without requiring labelled samples, supervised models are more promising as they normally achieve a higher classification accuracy on this task. Furthermore, we proposed a new augmentation approach, which shows advantageous in improving the performance of supervised learning models, e.g., logistic regression models and support vector machines.

References

[1] Min Chen, Shiwen Mao, and Yunhao Liu. Big data: A survey. *Mobile networks and applications*, 19(2):171–209, 2014.

[2] Brendon Hall. Facies classification using machine learning. *The Leading Edge*, 35(10):906–909, 2016.

[3] J Johnston, A Guichard, et al. New findings in drilling and wells using big data analytics. In *Offshore Technology Conference*. Offshore Technology Conference, 2015.

[4] Joel Gevirtz, Adriana Ovalle, et al. Lithofacies classification for earth modeling: Letting the data speak for itself. In *SPE Unconventional Resources Conference*. Society of Petroleum Engineers, 2017.

[5] Martin Hilbert. Big data for development: A review of promises and challenges. *Development Policy Review*, 34(1):135–174, 2016.

[6] Martin Wolf, Jacques Pelissier-Combescure, et al. Faciolog-automatic electrofacies determination. In *SPWLA 23rd Annual Logging Symposium*. Society of Petrophysicists and Well-Log Analysts, 1982.

[7] JM Busch, WG Fortney, LN Berry, et al. Determination of lithology from well logs by statistical analysis. *SPE formation evaluation*, 2(04):412–418, 1987.

[8] Jong-Se Lim, Joe M Kang, Jungwhan Kim, et al. Multivariate statistical analysis for automatic electrofacies determination from well log measurements. In *SPE Asia Pacific Oil and Gas Conference and Exhibition*. Society of Petroleum Engineers, 1997.

[9] Trond Mathisen, Sang Heon Lee, Akhil Datta-Gupta, et al. Improved permeability estimates in carbonate reservoirs using electrofacies characterization: a case study of the north robertson unit, west texas. In *SPE Permian Basin Oil and Gas Recovery Conference*. Society of Petroleum Engineers, 2001.

[10] Y Zee Ma. Lithofacies clustering using principal component analysis and neural network: applications to wireline logs. *Mathematical Geosciences*, 43(4):401–419, 2011.

[11] Yunxin Xie, Chenyang Zhu, Wen Zhou, Zhongdong Li, Xuan Liu, and Mei Tu. 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, 2018.

[12] Chun Che Fung, Kok Wai Wong, and Halit Eren. Modular artificial neural network for prediction of petrophysical properties from well log data. *IEEE transactions on instrumentation and measurement*, 46(6):1295–1299, 1997.

[13] Luisa Rolon, Shahab D Mohaghegh, Sam Ameri, Razi Gaskari, and Bret McDaniel. Using artificial neural networks to generate synthetic well logs. *Journal of Natural Gas Science and Engineering*, 1(4-5):118–133, 2009.

[14] Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

[15] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[16] Sotiris B Kotsiantis, I Zaharakis, and P Pintelas. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160:3–24, 2007.

[17] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152. ACM, 1992.

[18] Wentian Li, Jane E Cerise, Yaning Yang, and Henry Han. Application of t-sne to human genetic data. *Journal of bioinformatics and computational biology*, 15(04):1750017, 2017.

[19] Wahyu Caesarendra, Achmad Widodo, and Bo-Suk Yang. Application of relevance vector machine and logistic regression for machine degradation assessment. *Mechanical Systems and Signal Processing*, 24(4):1161–1171, 2010.

[20] Anil K Jain. Data clustering: 50 years beyond k-means. *Pattern recognition letters*, 31(8):651–666, 2010.