The effect of supervised feature extraction techniques on the facies classification using machine learning[1]

Hamid Reza Okhovvat[a], Mohammad Ali Riahi[b]*, Mohammad Mahdi Abedi[c]
[a]*Petroleum, Mining and Materials Engineering Department, Islamic Azad University, Central Tehran Branch.* h_okhovvat@yahoo.com
[b]*Institute of Geophysics, University of Tehran, Iran,* mariahi@ut.ac.ir
[c]*BCAM - Basque Center for Applied Mathematics, Spain.* mm.abedy@gmail.com

Abstract

The widely accepted supervised machine learning classification algorithms are used for the semi-automating of the feature extraction process. In the machine learning facies classification process, each wireline log is a feature in the feature space. Since features are important in classification decisions, using suitable features improves the performance of a classification algorithm.
In this study, three feature sets are compared containing the original conventional features (well-logs), and the extracted features from the unsupervised PCA and supervised FDA methods, using two classifier algorithms, namely SVM and RF. The FDA showed an improvement in the performance of facies classifiers while PCA can even deteriorate the results. An F1 score of 0.61 averaged over the available 20 folds for the combination of FDA feature extractor and RF classifier is achieved. This represents a 5% improvement in the prediction accuracy, compared to the conventional use of wells information as features with an F1 score of 0.56. Moreover, the conventional method uses all seven well-logs while with the FDA we only use three features.

Keywords: Facies classification, Fisher Discriminant Analysis (FDA), Machine learning, Principle Component Analysis (PCA), Random forest (RF), Support Vector Machine (SVM).

Introduction

Facies classification is a fundamental part of geologic investigations. It includes assigning a rock type or class to different measured properties. Accurate classification of facies and their distributions can provide a better insight into the depositional environment. Depositional facies provide useful information about rock properties such as permeability, porosity, density, and pore size making it possible to predict the variation of porosity and permeability in a reservoir.

A criterion for lithofacies classification is rock core samples extracted from drilled wells. Nevertheless, core samples may not be available. Due to the limited access to the core samples in comparison with the number of wells in the field, it is necessary to develop an efficient tool to classify lithofacies without cores (Dubois et al., 2006). A conventional petrophysical procedure to generalize lithofacies from cored wells to wells without core is matching the physical rock properties measured by wire-line logs in these wells. Manually assigning lithofacies is a cumbersome process that may be influenced by human errors because of the involvement of different interpreters.

[1] This is a preprint. This paper is published in Journal of Seismic Exploration Volume 31, Issue 6, Pages 563 – 577, December 2022.

The traditional linear methods are not always successful; mainly because: (a) Features (wire-line logs) are not linearly related to each other; (b) a feature space has overlap between different facies classes; (c) the increase in dimensions (increasing the number of well-logs) add to the complexity of the problem.

Today, benefiting from the fast development of artificial intelligence (AI) and the success of machine learning in solving non-linear classification problems, extensive studies have been done on implementing machine learning algorithms to rock facies classification. The earliest works were based on applying neural networks to rock-type classification (Wolf et al., 1982; Busch et al., 1987; Baldwin et al., 1990; Rogers et al., 1992; Kapur et al., 1998; Saggaf and Nebrija, 2000; Russell et al., 2002). Cuddy (1997) did a review of the implementation of fuzzy logic in petrophysics. Dubois et al. (2007) compared four machine-learning approaches to facies classification (Bayes', Fuzzy logic, K-Nearest Neighbor (KNN), and feed forward-back propagating artificial neural network). Most recently, there have been Support Vector Machine (SVM) (Wang et al., 2013; Hall, 2016) and ensemble method (Bestagini et al., 2017) implementations. Hall (2016) proposed a geophysical tutorial to demonstrate a basic implementation of machine learning techniques for facies classification. To improve the results of facies classification tasks, the Society of Exploration Geophysicists (SEG) held a machine learning contest (Hall, 2016), in which the participants were asked to use different methods to tackle this problem by utilizing various classifiers. The competition results are documented in Hall and Hall, 2017 most of them were based on the implementation of different classifier algorithms.

In this study, we first review the facies classification problem, followed by a description of the proposed facies classification algorithms. we improve the accuracy of the facies classification based on supervised feature extraction techniques. The strategy is based on adding a feature extraction step to the conventional workflow before classification (Figure 1). Using new extracted features that allow adequate separation of the facies classes instead of the original features, we improve the classification accuracy. We assess two different feature extraction methods, namely the supervised and unsupervised approaches. Finally, we apply our method to the 2016 SEG well-logging benchmark dataset and show how it further improves the reported best machine learning prediction results by Random Forest and Support Vector Machine classifiers.

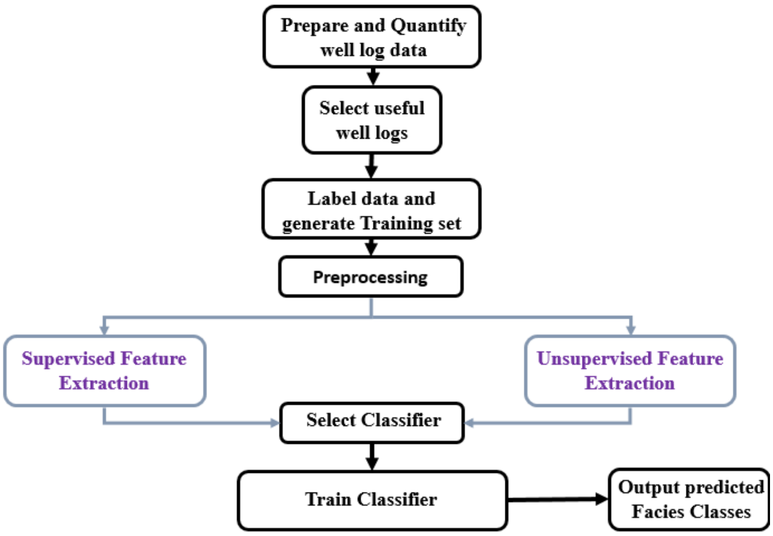
Prepare and Quantify well log data
Select useful well logs
Label data and generate Training set
Preprocessing
Supervised Feature Extraction
Unsupervised Feature Extraction
Select Classifier
Train Classifier
Output predicted Facies Classes

Figure 1: The workflow illustrates the steps involved in the supervised machine learning facies classification. Feature extraction step added to conventional works.

Methodology

Feature extraction

A feature plays a significant role in machine learning classification tasks. Features are extracted from an original set of measured data and are expected to be informative and non-redundant, assisting the subsequent classification (and learning) of problems. Therefore, in pattern recognition, it is more beneficial to insert a feature extraction step in the algorithm, before the classification. We utilize supervised and unsupervised feature extraction methods in facies classification which include feature extraction methods of Principal Component Analysis (PCA) and Fisher Discriminant Analysis (FDA).

Principal Component Analysis

There exist various techniques for dimensionality reduction; Principal Component Analysis (PCA) is one of the oldest and most commonly used ones. The PCA reduces a large set of primary features (conventional logs) to highlight variations in the data by a linear combination of original features and find new reduced features. PCA is a linear and unsupervised dimensionality reduction technique that projects data with higher dimensions into a lower dimension while preserving the main information in the original input data set.

Let m and n be the number of original variables and number of observations for each variable (i.e. the number of data samples), respectively. The input data set is then a matrix of dimension $\mathbf{X} \in \mathbb{R}^{n\times m}$, wherein $\mathbb{R}$ denotes the field of real numbers.
The standard PCA algorithm is briefly described as follows:

Step 1: Normalize the columns of **X** (variables) by subtracting the mean of each column followed by division with the standard deviation of each column in such a way that the mean and variance of each column are equal to 0 and 1, respectively.

Step 2: Compute covariance matrix **C:**

$$\mathbf{C} = \frac{1}{n-1}\mathbf{X}^T\mathbf{X}. \tag{1}$$

Step 3: Calculate the eigenvectors (**V**) and eigenvalues (**A**) from the covariance matrix (or correlation matrix) **C** by the eigenvalue decomposition (EVD):

$$\mathbf{C} = \mathbf{VAV}^T, \tag{2}$$

where

$$\mathbf{A} = diag(\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m \geq 0),$$

where λ_i is the set of eigenvalues. Of note, eigenvectors are arranged according to their eigenvalues in descending order. In this step feature extraction by PCA is completed and the new features are extracted as:

$$\mathbf{T} = \mathbf{XV}, \tag{3}$$

where **T** is the $n \times n$ transformed data in the PCA domain.

Step 4: Decision on the number of eigenvectors. Determining the number of principal components by another analysis method and decomposing **V** into a score space ($\mathbf{V}_{pc}$) and a residual space ($\mathbf{V}_{res}$):

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{pc} & 0 \\ 0 & \mathbf{A}_{res} \end{bmatrix}, \tag{4a}$$

$$\mathbf{V} = \begin{bmatrix} \mathbf{V}_{pc} \in \mathbb{R}^{m\times a} & \mathbf{V}_{res} \in \mathbb{R}^{m\times(m-a)} \end{bmatrix}, \tag{4b}$$

where,

$$\mathbf{A}_{\mathrm{pc}} = diag(\lambda_1.\lambda_2 \ldots.\lambda_a), \tag{5a}$$

$$\mathbf{A}_{\mathrm{res}} = diag(\lambda_{a+1}.\lambda_{a+2} \ldots.\lambda_m). \tag{5b}$$

Step 5: Compute the projection matrix **T**:

$$\mathbf{T} = \mathbf{X}\mathbf{V}_{pc}, \tag{6}$$

Where **T** is $n \times a$ reduced transformed data in the PCA domain.

Fisher discriminant analysis (FDA)

The process of PCA ignores class labels and does not consider the information (discrimination) between different classes during the calculation of the transformed matrix (**T**). This problem is addressed within Fisher Discriminant Analysis (FDA) technique. FDA transformation matrix includes vectors that maximize scatter between classes while minimizing the within-class separation. In other words, FDA can be considered, a supervised PCA.

Let x_i be a column vector constructed from the i*th* row of data set matrix, **X**. Besides, let *p* be the number of classes and n_j be the number of samples within the *j*th class.
The standard FDA framework is briefly formulated as follows:

Step 1: Compute the total-scatter matrix $\mathbf{S}_t$ by,

$$S_t = \sum_{i=1}^{n} (\mathbf{X}_i - \overline{\mathbf{X}})(\mathbf{X}_i - \overline{\mathbf{X}})^{\mathrm{T}}, \tag{7}$$

where $\overline{\mathbf{X}}$ represents the total mean vector.
Step 2: Compute intra-class scatter matrix $\mathbf{S}_w$

$$\mathbf{S}_w = \sum_{j=1}^{p} \mathbf{S}_j, \tag{8}$$

where:

$$\boldsymbol{S}_j = \sum_{X_i \in X_j}^{n} (\mathbf{X}_i - \overline{\mathbf{X}}_j)(\mathbf{X}_i - \overline{\mathbf{X}}_j)^{\mathrm{T}}, \tag{9}$$

also, $\overline{\mathbf{X}}_j$ denotes the mean of the *j*th class.

Step 3: Calculate the inter-class scatter matrix, $\mathbf{S}_b$:

$$\mathbf{S}_b = \sum_{j=1}^{p} n_j(\mathbf{X}_i - \overline{\mathbf{X}}_j)(\mathbf{X}_i - \overline{\mathbf{X}}_j)^{\mathrm{T}}, \tag{10}$$

we have

$$\mathbf{S}_t = \mathbf{S}_w + \mathbf{S}_b \ . \tag{11}$$

Step 4: The FDA vectors, **W**, can be obtained using the generalized eigenvalue decomposition:

$$\mathbf{S}_b\mathbf{W} = \lambda\mathbf{S}_w\mathbf{W}. \tag{12}$$

Since $rank(\mathbf{S}_b) < p$, there exist at most $p - 1$ eigenvectors corresponding to non-zero eigenvalues (Fukunaga, 1990). Simply put, the FDA can find at most $p - 1$ meaningful features (the remaining FDA features are arbitrary). This is a fundamental bottleneck of the FDA in dimensionality reduction. Let k be several non-zero eigenvalues :

$$\mathbf{W}_k = [\mathbf{w}_1 \quad \mathbf{w}_2 \quad \dots \quad \mathbf{w}_k] \tag{13}$$

Step 5: Calculate the FDA transformation vectors:

$$\mathbf{Z}_i = \mathbf{W}_k^T\mathbf{X}_i \tag{14}$$

Unlike PCA, FDA considers the between-classes information to compute projection vectors. Therefore, in multi-class classification problems, it is expected that FDA transformation vectors discriminate better between different classes compared to PCA (Fakhari and Hashemi, 2019).

Random forest (RF) and support vector machine (SVM) classifiers, which are reportedly robust (Hall and Hall, 2017; Mosser and Briceno, 2017; Chen and Guestrin, 2016), are chosen for machine learning facies classification. In the following, we give a brief description of them.

Random Forest

Random forest (RF) is based on the principle of using a decision tree as the basic classifier. Decision Tree is one of the easiest and most popular classification algorithms. It includes a series of nodes, a directional path that starts at the base with a single node and extends to the many leaf nodes that represent the categories that the tree can classify. Each node represents one of the features of our data, each branch explains a decision and each leaf shows a class label (Breiman, 2001).

Random Forest is an example of a learning methods ensemble that aggregates, or forest, multiple decision trees to limit overfitting and minimize error (Breiman, 2001). RF classifier outperforms most classifiers because of resistance to overfitting, ease of tuning parameters, and being fast (Kavzoglu, 2017). The final decision of determination of class label for a new sample is a decision made by combining individual tree votes in the decision forest. The main purpose of the RF classifier is to reduce error by creating multiple decision trees using a bootstrapped sampling method. The bootstrap method and the random feature selection can decrease the correlation between the classification trees by reducing the strength of individual trees, which reduces the risk of overfitting using uncorrelated trees (James et al. 2013; Pelletier et al. 2017).

Any RF classification model has two parameters that have to be tuned before model training: The number of trees in the forest and the maximum number of randomly selected features for node-splitting decisions (Cracknell and Reading 2014). We select these two parameters through cross-validation analysis. The best parameter pair is chosen by the lowest validation error.

Support Vector Machine

Support vector machine (SVM) is a supervised non-parametric classifier method that has been extensively used for classification and regression problems. The main idea of SVM is to find the optimal separating boundary which maximizes the margin of the two classes (Corres and Vapnik, 1995). For a detailed explanation of SVM, the theory refers to Corres and Varpnik (1995). The generalized SVM can classify nonlinear and multi-class data (Liu and Zheng 2005). In general, the SVM is a linear classifier. In this paper, we use a Radial Basis Function (RBF) kernel for the SVM classifier which is successfully used for non-linear classification models. The radial basis function (RBF) kernel is a popular kernel function employed in various kernel-based learning schemes.

SVM classifier has two hyperparameters that demand optimization: (i) The penalty function $\boldsymbol{C}$ which controls the trade-off between minimizing the training error and maximizing the classification margin and the model complexity; (ii) The gamma kernel parameters (γ). Thus, we need to find the best combination of $\boldsymbol{C}$ and γ to improve classification performance.

Experimental Results

Problem Dataset

SEG in 2016 provided a well-log dataset for their facies classification challenge. This dataset contains 3232 samples in 8 wells in the Hugoton field of southwest Kansas (Dubois et al., 2007). In this dataset, each sample was placed at a depth of one of the eight wells which comprised of five wireline log curves, two indicator variables, measured depth, and labeled facies data based on core analysis. Wireline log curves include gamma-ray (GR), resistivity (ILD_log10), photoelectric effect (PE), neutron density porosity difference (DeltaPHI), and average neutron density porosity (PHIND).

The cyclical vertical succession of the Council Grove Group reveals a pattern of eight main facies in this region. Puckette et al. (1995) derived a set of 9 lithofacies which are mostly used for lithofacies classification using machine learning in this area (Hall, 2016). The facies descriptions are labeled as Non-marine sandstone (SS), Non-marine coarse siltstone (CSiS), Non-marine fine siltstone (FSiS), Marine siltstone and shale (SiSh), Mudstone (MS), Wackstone (WS), Dolomite (D), Packstone grainstone (PS) and Phylloid-algal bafflestone (BS).

Figure 2 displays the wireline logs of an available well (Shankle) as well as interpreted lithofacies assigned in the well.

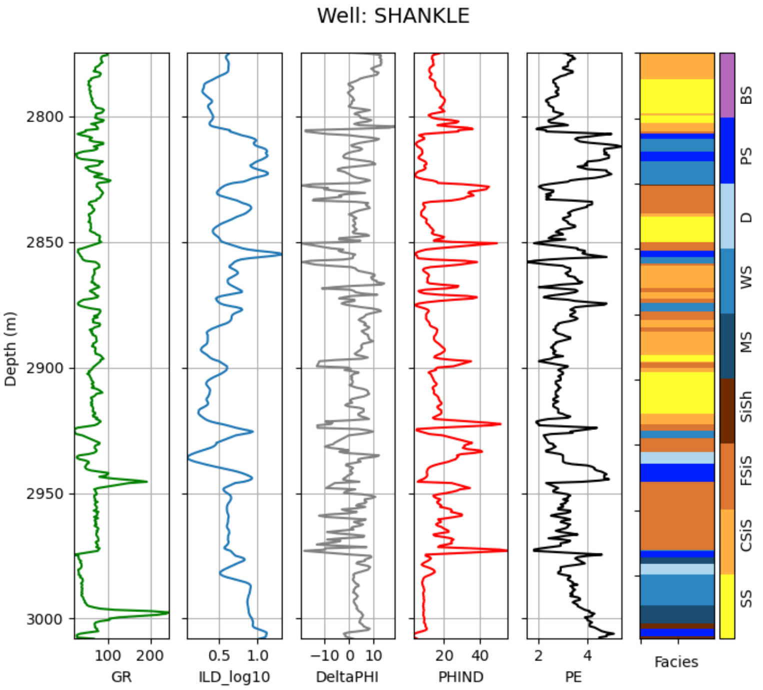

Figure 2: Displays the wireline logs: gamma-ray (GR), resistivity (ILD_log10), an average of the neutron and density log porosity (DeltaPHI), the difference between the neutron log and the density porosity (PHIND), photoelectric factor (PE) and interpreted lithofacies assigned in the Shankle well.

From a pattern recognition point of view, each wireline log information is called a 'feature'. In classification tasks, the feature selection criteria are based on the facies class separability. Figure 3 shows a cross-plot between wireline log parameters in which each facies category is indicated by a color. It can be seen that the facies (indicated by colors) are not completely separable in the feature space.

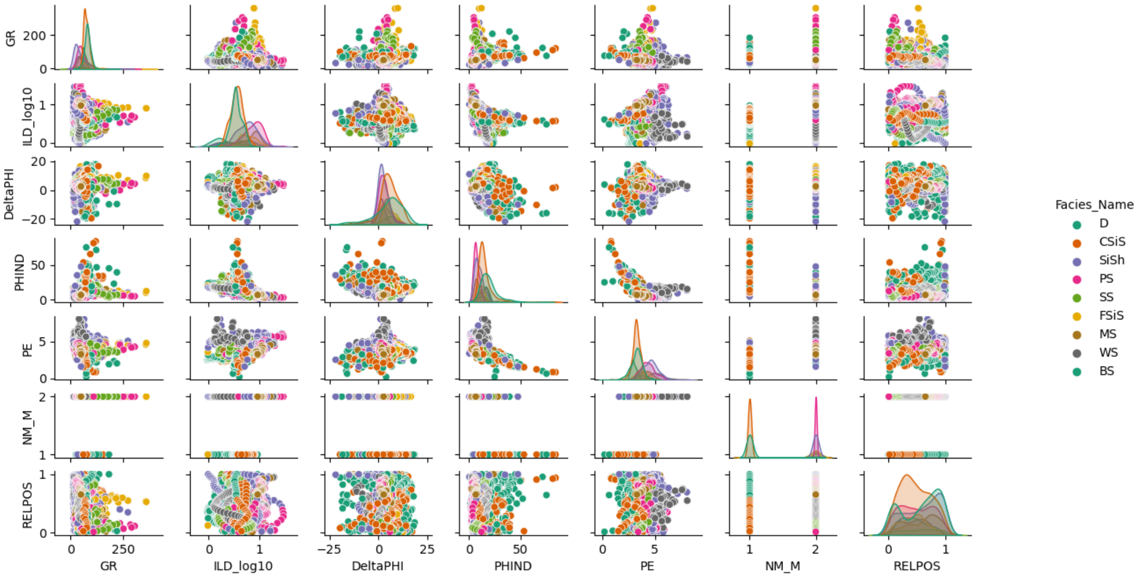

Figure 3: Scatter plot diagram of the distribution of wireline logs (PE, GR, ILD_log10, DeltaPHI, PHIND) is shown. Each facies is represented with different colors for better visualization.

To have a better facies classification, we propose to apply feature extraction methods, beforehand. This means generating new features from the available primary features for a better separation of the classes into extracted feature space.

Our experimental results have been carried out on published data sets for the facies classification described by Hall (2016). First, we implement both supervised and unsupervised feature extraction methods namely PCA and FDA on the original dataset before classification. We extract new features from a primary set of wireline logs to better separate facies classes in the new extracted space and further increase classification performance.

Figure 4 shows how facies classes are separated in the new space. In Figure 4 the cross-plot of three elements of log data (out of the seven primary features) is compared with the three first features extracted from FDA. In Figure 4a, the overlapping facies classes in 3D space are difficult to separate with parametrically certain boundaries in the logs domain. But in Figure 4b, we can observe that the facies classes are better clustered. Although FDA is not a clustering method, it can help visualize the patterns such as facies classes by reducing dimensionality. Due to a large number of facies classes (9 classes), these patterns might not be well visible on a 3D FDA plot, but they show up more clearly in higher-dimensional space.

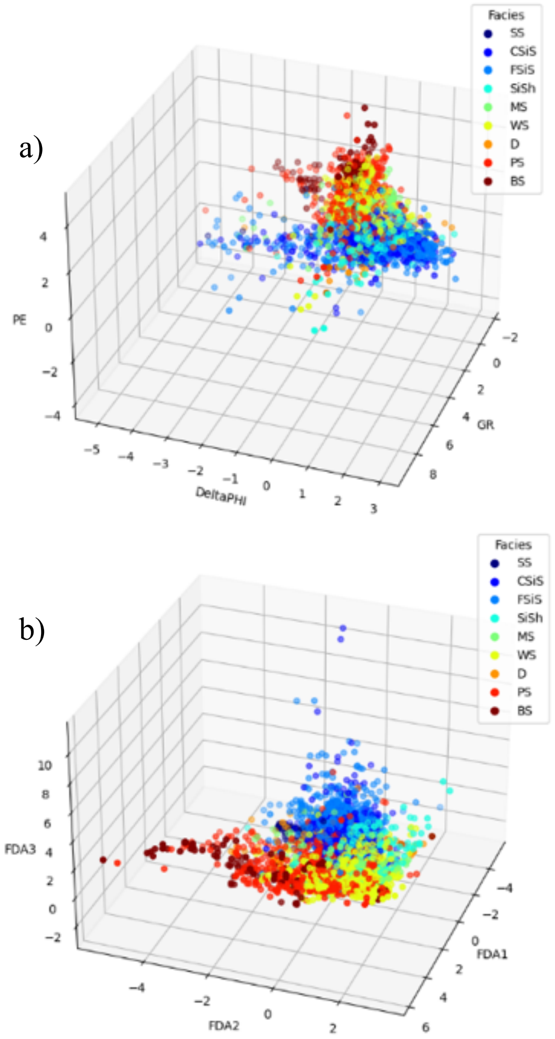

Figure 4: a) A representative cross-plot of three elements (out of the seven elements) of log data (raw features) compared with b) three first features extracted from FDA. Each facies class is depicted by color.

With the PCA and FDA extracted features we have three datasets: well logs, PCA features, and FDA features datasets to compare. To evaluate the predictive performance of a classification model, we need to split data into training and test parts. We select Shankle well data as the test well (similar to Hall, 2016) to evaluate the performance of the final classification process. We evaluate the classification models in terms of F1 score criteria to have a comparable assessment with those reported in Hall and Hall, 2017.

First, classification is performed via an SVM classifier and the conventional features (well logs) without using any feature extraction. In a similar test, Hall (2016) achieved an average F1 score of 0.43 on the blind test well (Shankle well). We obtained the average F1 score of 0.49 for the blind test well, using an SVM with Radial Basis Function (RBF) kernel which optimized its hyperparameters.

Next, we performed an SVM classifier using features extracted from PCA and FDA approaches.
To rank features in classification for every three datasets (well legs, PCA, FDA), we compare both SVM and RF classification methods by adding features in the learning step. For SVM and RF classifications, the average F1 score of 20 cross-validation experiments for each available dataset as inputs of classifier, are shown in Figures 5 and 6, respectively.

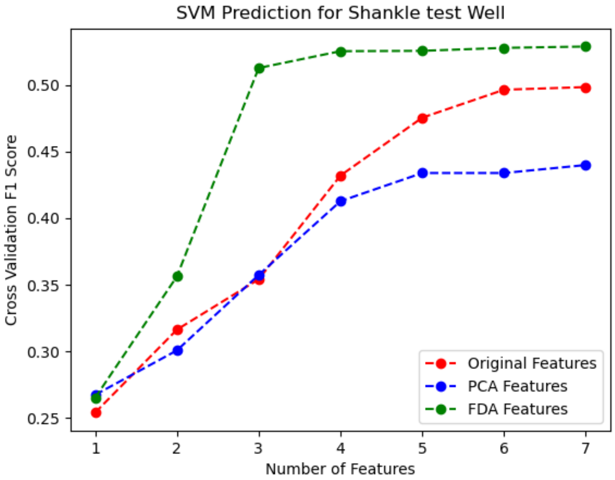

Figure 5: Averaged SVM classification F1 score after 20 cross-validation experiments for Original features (red), PCA (blue), and FDA (red) versus the number of features.

In Figure 5, the red line shows that the proposed SVM prediction accuracy for the test data (Shankle Well) increases by increasing the number of features in all three models. This trend indicates that more features provide more discriminative power. When using FDA features (green line in Figure 5), the F1 score grows fast with the addition of the first 3 FDA features and approximately stabilizes near a certain point. Another point of evidence in this figure is the failure of the unsupervised PCA method (blue line). Among the three input datasets provided by different features approaches, which are tested here, the FDA results in a higher cross-validation F1 score and converges to this score with fewer features.

Next, we repeat the classification problem by RF classifier. The RF predicts the facies classes with an F1 score of 0.56 for the Shankle blind well, when using the conventional features (well logs) without using any feature extraction. With the FDA-generated features, the F1 score improves up to 0.61 (Figure 5).

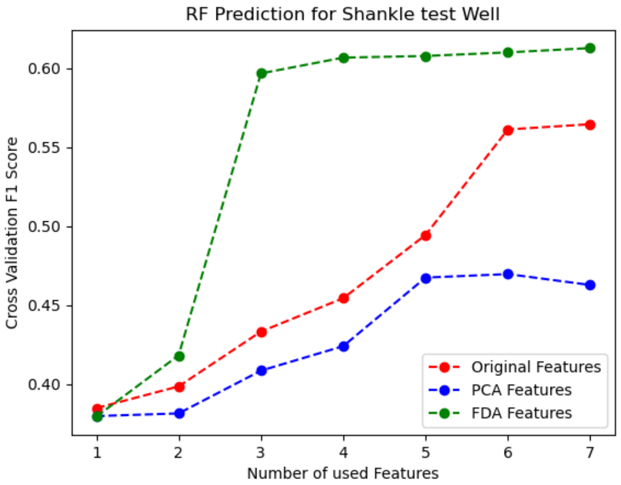

Figure 6: Averaged RF classification F1 score after 20 cross-validation experiments for Original features (red), PCA (blue), and FDA (red) versus the number of features.

Comparing the classification accuracy of the results of the two classifiers algorithms (Figures 5 and 6), the RF classifier performs better than SVM for all the available datasets (collected dataset with well logs, PCA, and FDA features). The accuracy of prediction for both classifiers is improved by inserting a supervised feature extraction (FDA) step before running classification. The FDA approach can also be useful for feature reduction purposes. In our problem, there are 9 facies classes and 7 original features. Since the number of facies classes is more than the number of primary original features, computed features from FDA have the same size as primary dimension space. However, as shown in Figures 5 and 6, acceptable results can be obtained by using only the first three FDA dimensions for both classifications by different algorithms.

A predicted facies label is obtained for each approach. Figure 7 shows a comparison between true facies labels, predicted facies from SVM, and RF classifiers with and without using the FDA approach on the test well.

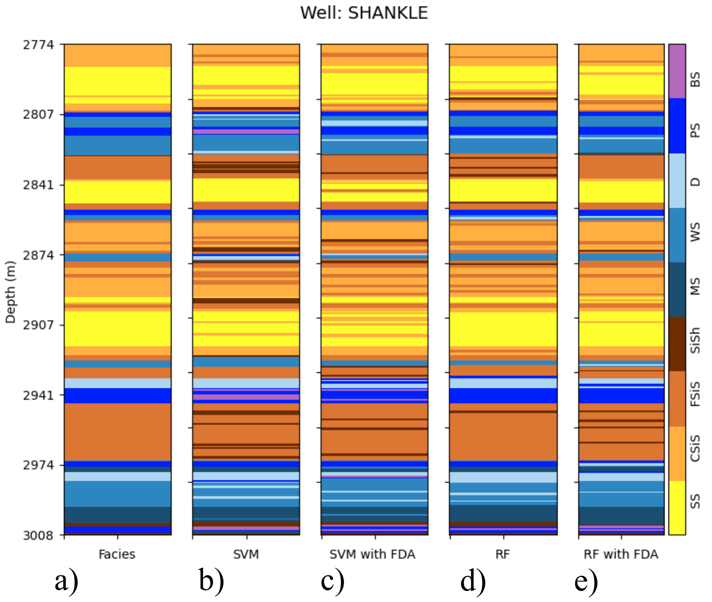

Figure 7: Representation of vertical facies sequence versus depth in the testing data from Shankle well for a) true facies, b) SVM classification results, c) combination of FDA with SVM, d) RF classification, and e) combination of FDA with FDA algorithm.

Figure 8 compares the predicted versus actual facies for predictions on the Shankle test data. In Figure 8 the five wireline logs are plotted as raw features, while the true and predicted facies labels are shown using the same depth information. We achieved the best predictions by combining the FDA approach with the RF classifier.

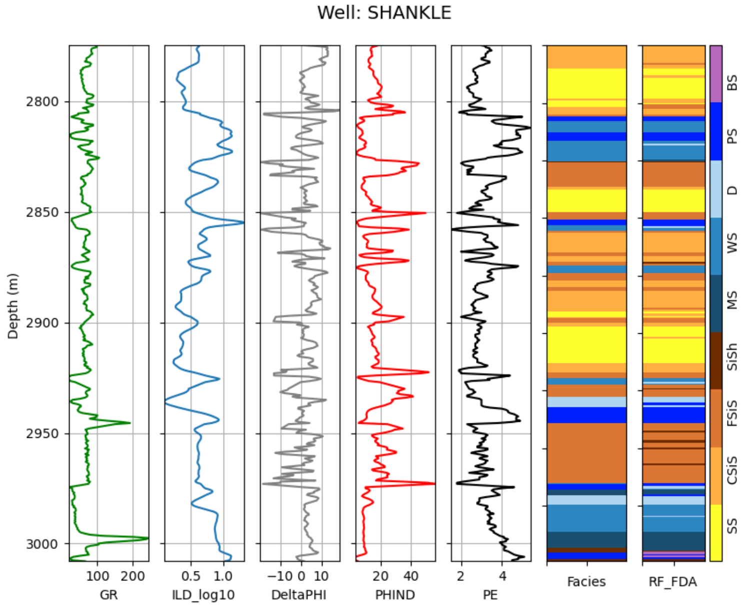

Figure 8: Wireline log measurements and comparison of facies classification between true labels and our proposed combination of FDA and RF methods.

Conclusions

The performance of facies classification based on conventional well-log is improved by implementing a supervised feature extraction method. To examine theverifiable 2016 SEG well dataset the Support Vector Machine (SVM) and Random Forest (RF) was applied. Then the results were compared with conventional well-log data as features versus extracted features as inputs of the machine learning classification process. Both supervised and unsupervised feature extraction techniques were implemented in facies classification. The obtained results from the Principal Component Analysis (PCA) as an unsupervised and Fisher Discriminant Analysis (FDA) as a supervised feature extraction method were compared. The supervised feature extraction method showed that it can improve machine learning facies classification results. The FDA improves the performance of facies classification and can reduce dimensions if a high number of conventional primary features are available. The proposed strategy provides a more accurate and consistent workflow for facies classification-based applications.

Declarations

Conflict of interest: The authors declare no competing interests.

References

Bestagini P., V. Lipari, and S. Tubaro, 2017, A machine learning approach to facies classification using well logs: SEG Technical Program Expanded Abstracts, 2137-2142.

Breiman L., 2001, Random forests. Mach Learn 45:5–32.

Mosser P., and A. Briceno, 2017, https://github.com/seg/2016-ml-contest/tree/master/LA Team.

Chen T.Q. and C. Guestrin, 2016, XGBoost: A Scalable Tree Boosting System. arXiv:1603.02754v3.

Cheung Y.M., and H. Jia, 2012, Unsupervised feature selection with feature clustering. In Proceedings of The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology-Volume 01. IEEE Computer Society. Pages 9–15.

Dubois M.K., G.C. Bohling, and S. Chakrabarti, 2007, Comparison of four approaches to a rock facies classification problem. Comput Geosci 33(5): 599–617.

Dubois M.K., A.P. Byrnes, G.C. Bohling, and J.H. Doveton, 2006, Multiscale geologic and petrophysical modelling of the giant Hugoton gas field (Permian), Kansas and Oklahoma, USA.

Dubois M.K., A.P. Byrnes, G.C. Bohling, S.C. Seals, and J.H. Doveton, 2003, Statistically-based lithofacies predictions for 3-D reservoir modeling: examples from the Panoma (Council Grove) Field, Hugoton Embayment, Southwest Kansas. In: Proceedings of the American Association of Petroleum Geologists annual convention, **12**, A44.

Fakhari M. G., and H. Hashemi, 2019, Fisher Discriminant Analysis (FDA), a supervised feature reduction method in seismic object detection. Geopersia, **9**, 141-149.

Friedman J. H., 2001, Greedy function approximation: A gradient boosting machine: The Annuals of Statistics, **29**, 1189-1232.

Fukunaga K., 1990, Introduction to Statistical Pattern Recognition. Academic Press, London.

Guyon I., and A. Elisseeff, 2003, An introduction to variable and feature selection. Journal of Machine Learning Research 3(Mar), 1157–1182.

Hall B., 2016, Facies classification using machine learning. Lead Edge **35**, 906–909.

Hall M., and B. Hall, 2017, Distributed collaborative prediction: results of the machine learning contest. Lead Edge **36**, 267–269.

Hall B., 2016, https://github.com/seg/2016-ml-contest.

Jolliffe I.T., 1986, Principal Component Analysis. Springer, New York, NY.

Liu Y., and Y.F. Zheng, 2005, One-against-all multi-class SVM classification using reliability measures. In Proceedings. 2005 IEEE International Joint Conference on Neural Network, **2**, 849-854. IEEE.

Pedregosa F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, and M. Blondel, 2011, Scikit-learn: Machine learning in python: Journal of Machine Learning Research, **12**, 2825-2830.

Vapnik V., 1995, The nature of statistical learning theory. Springer, Berlin.

Zhang L., and C. Zhan, 2017, Machine learning in rock facies classification - an application of XGBoost: International Geophysics Conference, Qingdao, China, 1371-1374.