Anuário do Instituto de Geociências
Universidade Federal do Rio de Janeiro
https://revistas.ufrj.br/index.php/aigeo/

ISSN 0101-9759
e-ISSN 1982-3908

Automatic Lithofacies Classification with t-SNE and K-Nearest Neighbors Algorithm

Classificação Automática de Litofácies Utilizando os Algoritmos t-SNE e K-NN

Guilherme Loriato Potratz[1,2], Smith W. A. Canchumuni[1], Jose David Bermudez Castro[1], Júlia Potratz[1,3] & Marco Aurélio C. Pacheco[1]

[1] Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Engenharia Elétrica, Rio de Janeiro, RJ, Brasil
[2] Universidade do Estado do Rio de Janeiro, Programa de Pós-Graduação em Geociências, Rio de Janeiro, RJ, Brasil
[3] Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Engenharia Elétrica, Programa de Pós-Graduação em Métodos de Apoio à Decisão, Rio de Janeiro, RJ, Brasil
E-mails: geo.loriato@gmail.com; saraucoc@uni.pe; bermudezjosedavid@gmail.com; jupotratz@gmail.com; marco@ele.puc-rio.br

Abstract

One of the critical processes in the exploration of hydrocarbons is the identification and prediction of lithofacies that constitute the reservoir. One of the cheapest and most efficient ways to carry out that process is from the interpretation of well log data, which are often obtained continuously and in the majority of drilled wells. The main methodologies used to correlate log data to data obtained in well cores are based on statistical analyses, machine learning models and artificial neural networks. This study aims to test an algorithm of dimension reduction of data together with an unsupervised classification method of predicting lithofacies automatically. The performance of the methodology presented was compared to predictions made with artificial neural networks. We used the t-Distributed Stochastic Neighbor Embedding (t-SNE) as an algorithm for mapping the wells logging data in a smaller feature space. Then, the predictions of facies are performed using a k-nearest neighbors (K-NN) algorithm. The method is assessed in the public dataset of the Hugoton and Panoma fields. Prediction of facies through traditional artificial neural networks obtained an accuracy of 69%, where facies predicted through the t-SNE+K-NN algorithm obtained an accuracy of 79%. Considering the nature of the data, which have high dimensionality and are not linearly correlated, the efficiency of t SNE+KNN can be explained by the ability of the algorithm to identify hidden patterns in a fuzzy boundary in data set. It is important to stress that the application of machine learning algorithms offers relevant benefits to the hydrocarbon exploration sector, such as identifying hidden patterns in high-dimensional datasets, searching for complex and non-linear relationships, and avoiding the need for a preliminary definition of mathematic relations among the model's input data.

Keywords: Facies prediction; Well logging; t-SNE

Resumo

Um dos processos críticos na exploração de hidrocarbonetos é a identificação e previsão das litofácies que compõem o reservatório. Uma das maneiras mais eficientes e mais baratas de realizar esse processo é a partir da interpretação de dados de perfilagem de poço, que são frequentemente obtidos de forma contínua e na maioria dos poços perfurados. As principais metodologias utilizadas para correlacionar dados de perfilagem aos dados obtidos em testemunhos de poço são baseados em análises estatísticas, modelos de machine learning e redes neurais artificias. O objetivo deste trabalho foi testar um algoritmo de redução de dimensionalidade de dados em conjunto com um método de classificação não supervisionado para predição automática de litofacies. O desempenho da metodologia apresentada foi comparado com predições feitas com redes neurais artificiais. Utilizamos a Incorporação Estocástica de Vizinho Distribuída (t-SNE) como um algoritmo para mapear os dados de registro de poços em um espaço menor. Em seguida, as previsões de fácies são realizadas usando um algoritmo K-NN. O método é avaliado no conjunto de dados público dos campos Hugoton e Panoma. A previsão de fácies através de redes neurais artificiais tradicionais obteve uma precisão de 69%, enquanto as fácies previstas através do algoritmo t-SNE + K-NN obtiveram uma precisão de 79%. Considerando a natureza dos dados, que têm alta dimensionalidade e não são linearmente correlacionados, a eficiência de t-SNE+K-NN pode ser explicada pela capacidade do algoritmo de identificar padrões ocultos em um limite nebuloso no conjunto de dados. É importante ressaltar que a aplicação de algoritmos de machine learnign apresenta benefícios significativos para o setor de exploração de hidrocarbonetos, dentre eles a identificação de padrões ocultos em um conjunto de dados de alta dimensionalidade, a busca de relações complexas e não-lineares, além de evitar a necessidade da definição prévia das relações matemáticas entre os dados de entrada do modelo.

Palavras-chave: Predição de fácies; Perfilagem de poços; t-SNE

1 Introduction

Challenges related to the intense demand for natural resources such as hydrocarbons, minerals, and groundwater, makes it increasingly necessary to have a detailed understanding of subsurface geology. This knowledge is essential to develop geological models that serve as support for exploration and exploitation projects of these resources in a sustainable and economically viable manner. The most efficient way of recognizing lithological sequences in the subsurface is through the description of well-cores. However, the core collection process is considerably expensive and not always having coverage of the entire study area (Cunha *et al*., 2003; Albuquerque *et al*., 2005; Rosa *et al*., 2008).

Regarding the high cost of collecting cores throughout the exploration area, geophysical profiling has proved to be a viable tool for gathering information about sequences of rocks from the subsurface (Burke *et al*., 1969; Delfiner *et al*., 1987). The above is possible because the measurements are practically continuous in the well, and the response of the physical properties of the profiles provides a close approximation of the rocks present in the well. According to Dubois *et al*. (2007), the classification of different types of rocks based on geophysical profiling data is fundamental for geological researches.

Notwithstanding presenting a lower cost, the measurements acquired by profiling tools not only represent the variations in lithology but also express changes in the medium's physical properties. Besides, for each class or intervals of continuous facies, there is a wide range of responses for each measured property, so that such responses can overlap information related to different facies. Then, this uncertainty in the measurement makes a challenge the characterization of the rock based just on geophysical profiling data.

An alternative has been attempting to establish correlations between data obtained from geophysical profiling and information from wells-cores. The main idea is to take advantage of both approaches to get a better characterization of rocks. Accordingly, several authors have tested traditional machine learning algorithms for this purpose (Busch *et al*., 1987; Rogers *et al*., 1992; Hsieh *et al*., 2005; Dubois *et al*., 2007), among others. However, these methods required wells-core samples for each corresponding profile sample which make the process costly. For instance, Dubois *et al*. (2007) used an artificial neural network for classifying facies based on profiling data. In this method, the well-core information was given indirectly from the facies label information. Results were satisfactory in comparison with traditional statistical methods and other machine learning algorithms, for instance, classical parametric methods using Bayes' rule, k-nearest neighbor, fuzzy logic, among others.

In this work, we propose a new semi-supervised method that requires a few sets of labeled profiling samples. We present a hybrid method based on the t-Distributed Stochastic Neighbor Embedding (t-SNE) (Maaten & Hinton, 2008) and K-Nearest Neighbors (K-NN) (Piegl & Tiller, 2002) methods which represents a more realistic scenario where wells-cores samples are usually scarce. The experiments were carried out in a public dataset, and the performance of the models was compared with a state of the art totally supervised method.

2 Fundamentals

2.1 t-Distributed Stochastic Neighbor Embedding

t-Distributed Stochastic Neighbor Embedding (t-SNE) is a probabilistic technique introduced by Maaten & Hinton (2008) for visualizing high dimensional data into the 2D or 3D dimensional space, attempting to preserve as much as possible the local structure of the data in the low-dimensional space. Although it was initially proposed for data visualization, it has been extended as a technique for the clustering of high dimensional data in any Euclidean space (Shahan & Steinerberger, 2017; Aibar *et al*., 2017; Linderman & Steinenerberger, 2019).

The t-SNE transforms the Euclidian distances between datapoints, at both the high-dimensional space and defined low-dimensional space, into conditional probabilities, such that nearby datapoints present high conditional probabilities, whereas, for widely separated datapoints, they are close to zero. Let x_i and x_j indicate two high-dimensional datapoints with conditional probability $p_{j|i}$, and y_i and y_j the associated mapped low-dimensional datapoints with conditional probability $q_{j|i}$, the t-SNE algorithm is trained to reduce the difference between $p_{j|i}$ and $q_{j|i}$. Mathematically, the conditional probabilities $p_{j|i}$ and $q_{j|i}$ for each datapoint are defined as follow, where σ_i represents the variance of the Gaussian.

$$p_{j|i} = \frac{exp\left(-||x_i - x_j||^2/2\sigma_i^2\right)}{\sum_{k\neq i} exp\left(-||x_i - x_k||^2/2\sigma_i^2\right)} \quad (1)$$

$$q_{j|i} = \frac{exp\left(-||y_i - y_j||^2\right)}{\sum_{k\neq i} exp\left(-||y_i - y_k||^2\right)} \quad (2)$$

The distance between $p_{j|i}$ and $q_{j|i}$ is reduced by minimizing the sum of Kullback-Leibler (KL) divergences over all datapoints using a gradient descent algorithm. Then, the cost function is given by

$$L_{t-SNE} = \min KL(P||Q) = \sum_i KL(P_i||Q_i) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{p_{j|i}} \quad (3)$$

2.2 Artificial Neural Network

An Artificial Neural Network (ANN) is a mathematical or computational model for processing information that mimics the way the biological nervous systems process information. As illustrated in Figure 1, it is a directed graph composed of processing units (neurons), organized by layers and interconnected layer-wise. These interconnections correspond to the weights, also known as the network's parameters, that are adjusted during training by minimizing a specific lost cost function using the back-propagation algorithm. A regular ANN is made by an input, a hidden or a stack of hidden layers, and an output layer. For multi-class classification is usually preferred the cross-entropy as the cost function, while for regression is selected the mean square error.

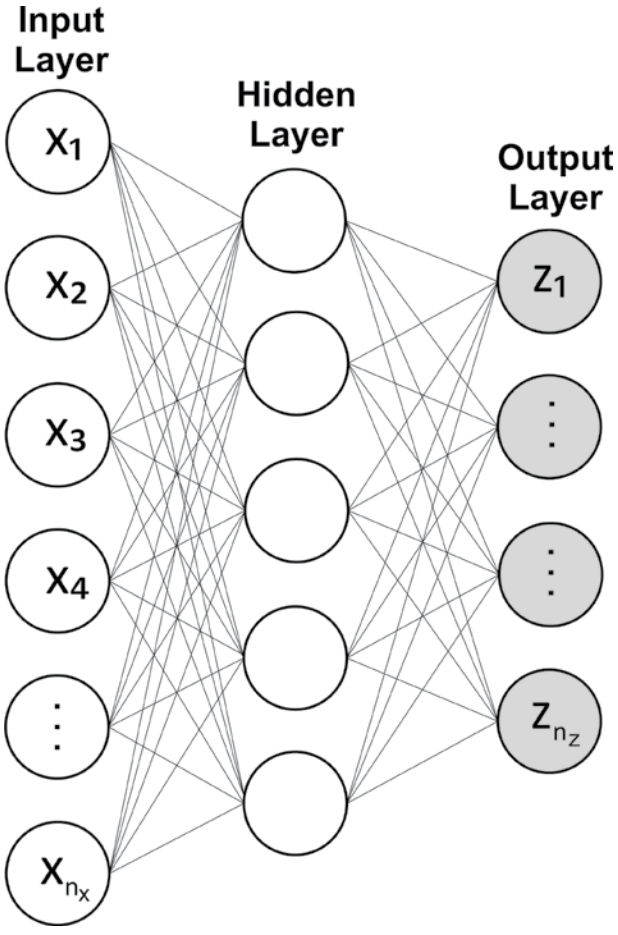

Figure 1 Base multilayer neural network architecture.

Formally, let´s denote the input, hidden and output layers as the x, h, and z vectors, respectively. The information fed to the ANN is forward processed layer-wise, as follows.

$$h = \sigma(W_i x + b_i) \quad (4)$$

$$z = \gamma(W_0 h + b_0) \quad (5)$$

Where **W***i* and **b***i*, and **W***o* and **b***o* denote the input and output weight matrices and bias vectors, respectively, and σ (*) and γ (*) the activation functions, which usually are non-linear, such as sigmoid, softmax, ReLu or others.

3 Methodology

In this work, we propose the use of a semi supervised methodology for classifying facies based on profiling data and the well-core information provided by the labeled classes. Particularly, we use a hybrid method that combines the capability of the t-SNE algorithm for dimensionality reduction and clusterization in conjunction with the K-NN classification algorithm, which usually requires a few sets of labeled samples to operate.

The adopted methodology is divided into two stages: unsupervised learning and supervised classification, as illustrated on the right side of Figure 2. During the unsupervised phase, the t-SNE algorithm learns to map the input feature space to another feature space where samples are grouped based on the Kullback-Leibler similarity metric. Then, the K-NN classifies the non-labeled samples in the new feature space according to the minimum Euclidian distance to the labeled samples.

As illustrated on the left side of Figure 2, before applying the t-SNE, the samples are almost grouped in just one cluster so that it is difficult to be discriminated using a linear classifier. Then, after applying the t-SNE, it can be observed well defined clusters that can be easily discerned by a simple linear classifier as K-NN method.

In this scheme, we first organize the samples per well and order them by the well-depth. Then, following the sliding windows procedure with overlapping, we sampled across the well-logging variables to create the features vectors. The objective of this process is to introduce correlations between the well-depth and the measured physical properties. Therefore, the samples input data vector for the t-SNE algorithm is formed by sequences of well-logging variables, as described in Equation 6.

$$data_{input} = [well\_log^{1}_{seq}, well\_log^{2}_{seq}, \dots, well\_log^{m}_{seq}] \quad (6)$$

where the subscript seq indicates the set of continuous measures according to the depth of the well, being **n** the window size used for sampling.

$$well\_log^{*}_{seq} = [well\_log^{*}_{-n}, well\_log^{*}_{-n+1}, \dots, well\_log^{*}_{1}] \quad (7)$$

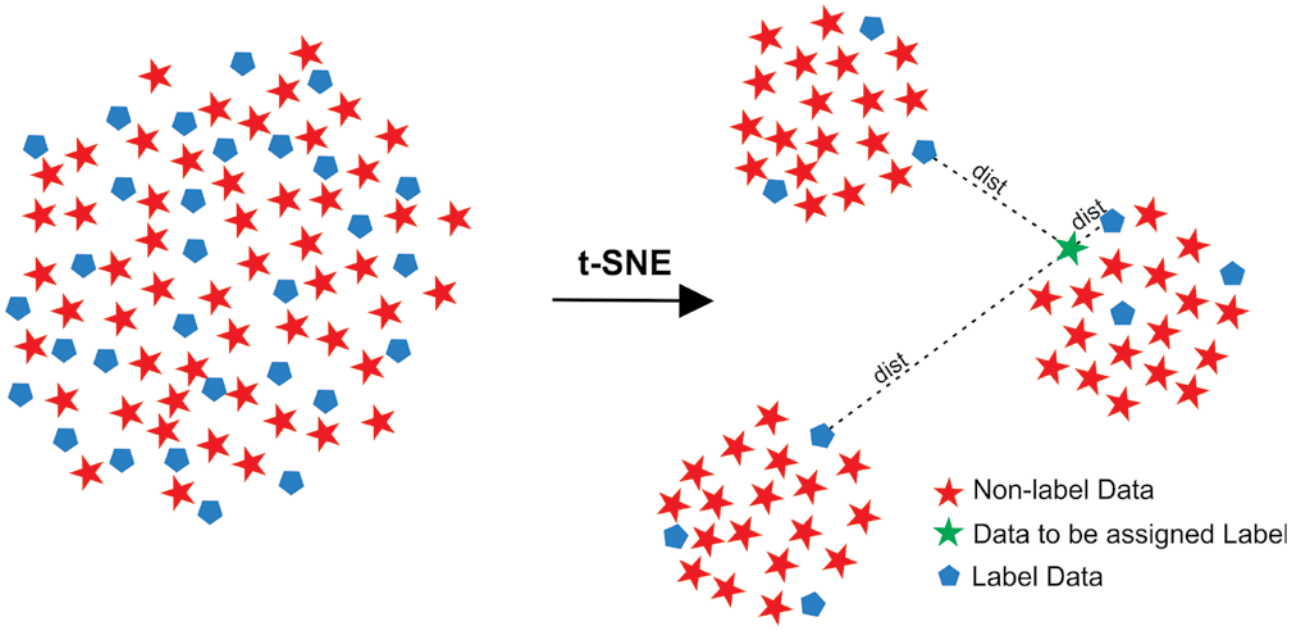

Figure 2 Methodology adopted for classification of geological facies. The stars and pentagons represent non-labeled and labeled samples, respectively. For a given non-label sample (green star) is computed the Euclidean distance regarding each labeled sample, and the labeled is assigned according to the samples with lower distance (K-NN method).

4 Experiments

4.1 Dataset

The dataset refers to the Hugoton and Panoma gas fields, in Kansas - USA (Dubois *et al*., 2003; Dubois *et al.,* 2006; Dubois *et al*., 2007). Those fields are located in the Anadarko Basin, bordered by the arch of Las Animas and by Kansas' central elevation, a Foreland-type basin associated with the beginning of the orogenesis of the Pensylvanian Ouchita-Marathon (Kluth, 1986; Perry, 1989). The main reservoirs are carbonate rocks, but, secondarily, there are sandstone reservoirs with high permeability and porosity. Sealing rocks are mostly very fine to coarse silstones and evaporites (Heyer, 1999; Dubois *et al*., 2003).

The Hugoton and Panoma fields are concentrated in the Chase and Council Grove groups, which represent vertical successions of lithofacies of well-known cyclic nature. Facies successions present a rising pattern, resulting from depositional environments controlled by quick floatation at relative sea level (Olson *et al*., 1997). Details concerning to the depositional model attributed to the Chase and Council Grove groups cam be found in the study by Dubois *el al*. (2006) and references therein.

Data related to the Hugoton and Panoma fields, used in this study, were provided by the University of Kansas and obtained with the challenge of predicting lithologies organised by the Society of Exploration Geophysicists. The same dataset was used in the studies of Dubois *et al*. (2003, 2006, 2007), and those data were used in their raw form, without any quality control.

This dataset contains a total of 3232 records from eight wells (SHRIMPLIN, SHANKLE, LUKE G U, CROOS H CATTLE, NOLAN, Recruit F9, NEWBY and CHURCHMAN BIBLE), corresponding to measurements at

0.15 m intervals of gamma rays (GR), resistivity (ILD_log$_{10}$), mean porosity and neutron density (PHIND). Additionally, it is included the difference between porosity and neutron density (DeltaPHI), and the information corresponding to the relative position (RELPOS). Global statistics of this dataset are summarized in Table 1, which lists the mean, standard deviation, min, and max values of the well-logging variables (GR, ILD_log$_{10}$, DeltaPHI, and PHIND).

Likewise, in Table 2 is exhibited the description of the geological facies codes that can be found in the wells, as well as the number of samples per facies. Notice that the dataset is completely imbalanced, being CSIS and FSIS the most representative classes whereas D is the less representative.

Additionally, in Figure 3 is illustrated an arrange of sixteen plots that describe the distribution of samples for each facies concerning the well-logging variables. Specifically, it is presented in the main diagonal of the arrange, the distribution of samples for each well-logging variable regarding each facies, while in the off-diagonal locations, there are presented scatter plots of bivariate diagrams of well-logging variables for each facies. Notice that most of the facies samples overlap between them making difficult the classification of the samples in the original feature space. This analysis can also be performed considering the boxplot diagram shows in Figure 4. It can be observed the overlapping of facies samples regarding the well-logging variables.

4.2 Experimental Setup

We first split randomly the dataset into two sets, one set for training and another set for testing, using a proportion of 2/3 and 1/3, respectively. Figure 5 shows two sets of bar diagrams associated with the distribution of samples per facies for the training and testing set, respectively. The distribution of samples regarding the facies is similar for both sets.

For improving the convergence of the assessed models, we normalized the well-logging variables to zero-mean and unit variance. Then, we balanced the dataset downsampling the most representative classes and replicating the less representative for training. This process was performed to avoid the models to be biased for the most representative classes.

Besides, to evaluate the capacity of the t-SNE+K-NN approach against the number of training labeled samples, we decided to consider only 50% of the training samples during the classification stage performed by the K-NN algorithm. For the t-SNE, we used the scikit-learn implementation, setting the learning rate to 120, the batch size to 32, and the training was stopped when no improvements were observed

Table 1 Statistic distribution of well logging variables used for automatic faces classification.

	GR	ILD_log$_{10}$	DeltaPHI	PHIND	NM_M	RELPOS
Count	3232	3232	3232	3232	3232	3232
Mean	66.14	0.64	3.55	13.48	1.5	0.52
Std	30.85	0.24	5.23	7.7	0.5	0.29
Min	13.25	-0.03	-21.83	0.55	1	0.01
Max	361.15	1.48	18.6	84.4	2	1

Table 2 Description and geological facies code.

Code	Description	Facies	Samples
1	Nonmarine sandstone	SS	259
2	Nonmarine corse siltstone	CSiS	738
3	Nonmarine fine siltstone	FSiS	615
4	Marine siltstone and shale	SiSh	184
5	Mudstone	MS	217
6	Wackestone	WS	462
7	Dolomite	D	98
8	Packstone-grainstone	PS	498
9	Phylloid-algal bafflestone	BS	161

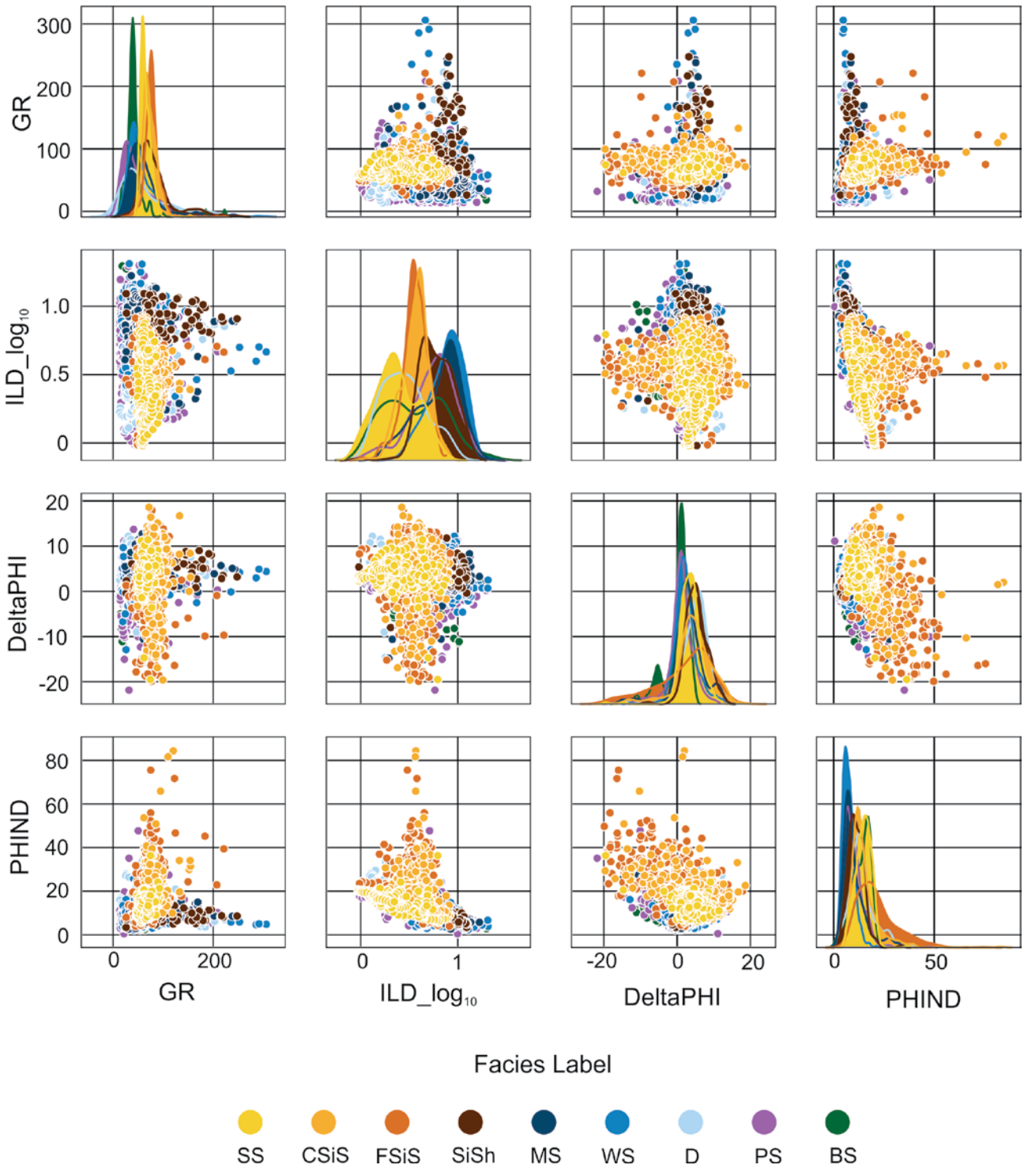

Figure 3 Bivariate diagrams for evaluating the behavior of the measured properties (GR, ILD_log$_{10}$, DeltaPHI and PHIND) in each of the nine facies. The facies are represented by colored dots that form a cloud where the boundaries between facies cannot be defined due to overlapping data.

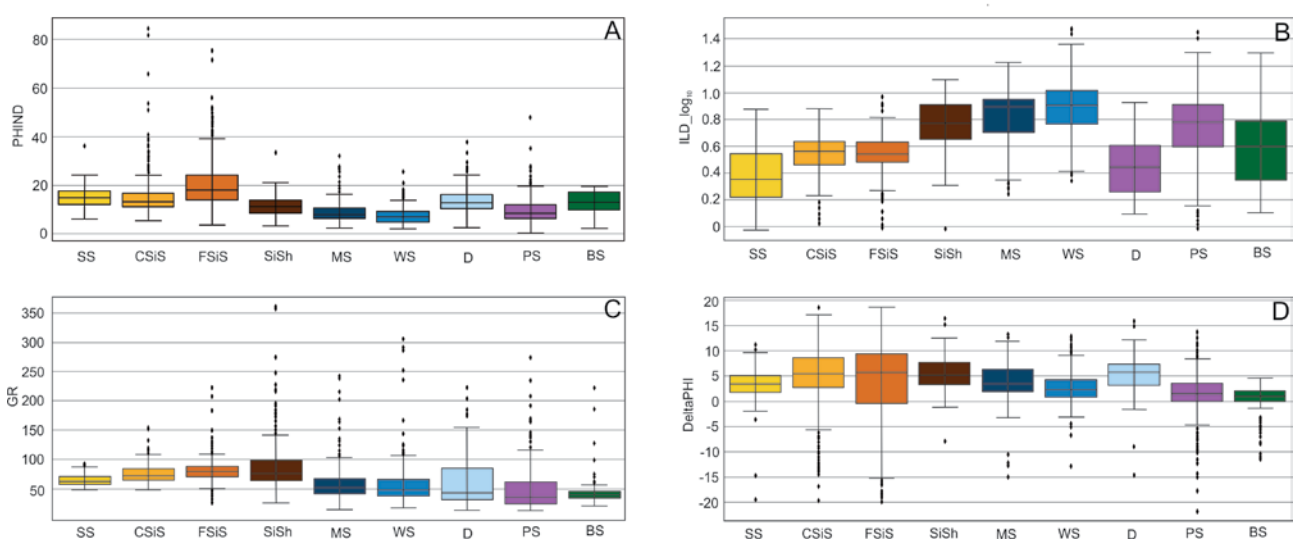

Figure 4 BoxPlot diagrams in which it can be observed the overlap in the space of the measured properties for each of the facies; A. PHIND; B. ILD_log$_{10}$; C. GR; D. DeltaPHI. Facies are represented by colored rectangles that show the variation of the measured data and the points scattered by the diagrams represent the outliers in each variation.

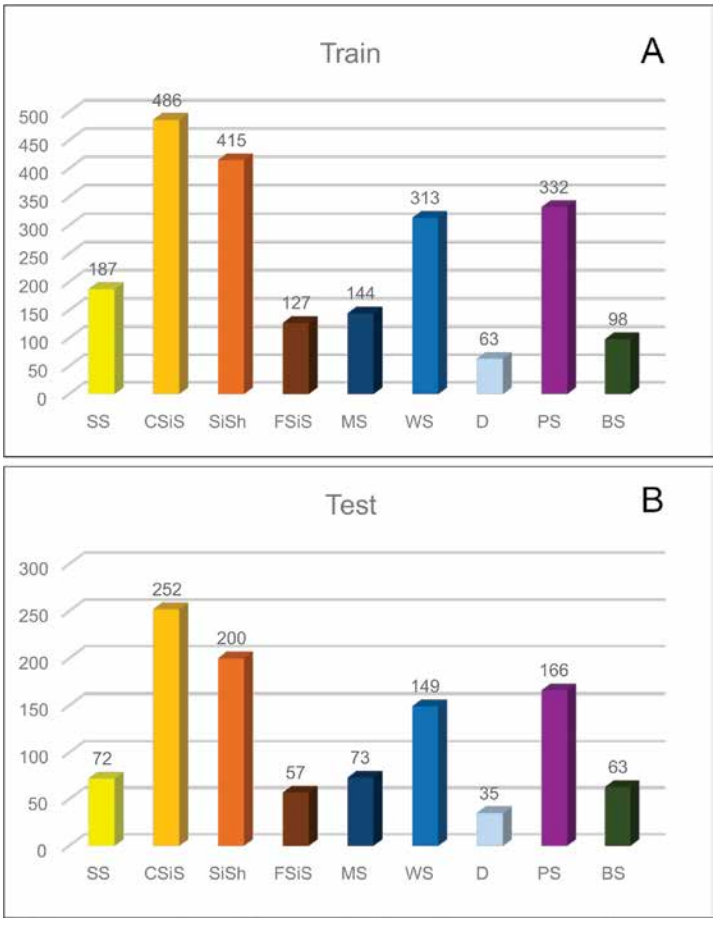

Figure 5 Separation of training and test sets at random, maintaining the proportionality of the samples; A. The number of samples from each facies in the training set is shown; B. The number of samples in the test set is shown.

in the KL lost, and for the K-NN, the K parameter was set to two. The repository containing codes in Python used to build this methodology are publicly available and can be accessed at: https://github.com/geoloriato/AUTOMATIC-FACIES-CLASSIFICATION-WITH-T-SNE-AND-K-NEAREST-NEIGHBORS-ALGORITHM

Regarding the ANN, we used the same network architecture presented in by [6]. Here, the number of hidden layers was set to 50 neurons, and the output layer to nine, corresponding to the equal number of classes. The experiments were carried out in the Keras framework, setting the batch size to 32, learning rate to 0.1, stochastic gradient descent (SGD) as optimization algorithm, and the number of epochs to 100.

The models were assessed in terms of Precision, Recall, and F1-score performance metrics for each class and average. Additionally, we also present a classification visual inspection analysis for two selected wells. Finally, the performance of the methods is evaluated considering the genesis of the rocks. We divide facies into four groups like a (Dubois *et al*., 2007):

- Group 1: sediments of non-marine origin (SS, CsiS, and FsiS);
- Group 2: clay of marine origin (SiSh);
- Group 3: chemical / physical carbonate rocks (MS, WS, D, and PS); and
- Group 4: corals (BS).

5 Results and Discussions

First of all, it is important to stress that the data available online needs to be used with caution. In this study, we use data with high degree of trustworthiness, which have been used in previous studies, such as Dubois *et al*. (2003, 2006, 2007), besides having been validated by the Society of Exploration Geophysicists. It is also relevant to stress that data of that nature are seldom made available to the public for free, which makes it even more difficult for studies like this to be developed.

Tables 3 and Table 4 summarize the performance obtained by the proposed method and the baseline, respectively, in terms of confusion matrix. Besides, in the same tables, it is also reported the precision, recall, and F1-score performance metrics per facies class. By comparing the confusion matrices, it is observed that the t-SNE+K-NN approach classified correctly more samples than the ANN counterpart for all classes. Notice that the lower improvement occurred for the SiSh facies, from 39 to 42 samples, and the best for the CSiS, which passed from 145 to 202 samples classified correctly. Regarding the major misclassifications, they existed between CSiS and FSiS, and WS and PS facies samples for both assessed methods.

Considering the performance in terms of Averages precision, recall and F1-score, the t-SNE+K-NN method achieved an improvement of 10% approx. in comparison with the baseline that presented a classification rate close to 70%. This behavior is consistent with almost all individual

Table 3 Confusion matrix obtained for the t-SNE+K-NN approach. In addition, it also presented the precision, recall, and F1-Score performance metrics.

		Predict									
		SS	CSiS	FSiS	SiSh	MS	WS	D	PS	BS	Total
True	SS	67	4		1						72
	CSiS	7	202	39	1	1		1	1		252
	FSiS	3	36	158	1				2		200
	SiSh		1	1	42	2	5	1	5		57
	MS	1		1	4	50	3	3	11		73
	WS			2	8	4	113	1	21		149
	D			1			2	26	6		35
	PS			1	7	9	17	2	129	1	166
	BS					1	1	1	1	**59**	63
	Precision	0.86	0.83	0.78	0.66	0.75	0.80	074	0.73	0.98	0.79
	Recall	0.93	0.80	0.79	0.74	0.68	0.76	0.74	0.78	0.94	0.79
	F1	0.89	0.82	0.78	0.69	0.71	0.78	0.74	0.75	0.96	0.79

facies. For instance, the lower precision rate for the ANN was 56 (MS) while for the t-SNE+K-NN was 66 (SiSh); the remaining facies got precisions scores from 73 (PS) up to 98 (BS). In terms of recall, the differences in performance were more notorious, presenting the t-SNE+K-NN recall rates from 68 (MS) up to 94 (BS), while the ANN was from 36 (MS) up to 91 (BS). The unique facies where the ANN obtained better results correspond to the SiSH class, where this method was superior in just 2%.

Figure 6 shows a graph that contrasts the performance of the models in terms of F1-Score. The blue and orange lines indicate the t-SNE+K-NN and ANN methods, respectively. It is observed that the higher improvement scored for the proposed methodology occurred for the MS facies, going from 44 up to 71. In contrast, for the SiSH, both methods performed similarly, 69 and 70. Regarding the performance over the other facies, it can be discerned

Table 4 Confusion matrix obtained for the ANN approach. In addition, it also presented the precision, recall, and F1-Score performance metrics.

		Predict									
		SS	CSiS	FSiS	SiSh	MS	WS	D	PS	BS	Total
True	SS	57	22	2							81
	CSiS	16	145	35	2				1		199
	FSiS	6	28	120		1	1		5		161
	SiSh			2	39	3	5		2		51
	MS	2	4		4	23	18	1	12		64
	WS			4	10	7	84	1	32		138
	D			1	3	1	2	18	3	3	31
	PS		3	7	3	6	13	2	98	2	134
	BS						1	1	3	48	53
	Precision	0.70	0.72	0.70	0.64	0.56	0.68	0.78	0.63	0.91	0.69
	Recall	0.70	0.73	0.75	0.76	0.36	0.61	0.58	0.73	0.91	0.69
	F1	0.70	0.72	0.72	0.70	0.44	0.64	0.67	0.68	0.91	0.69

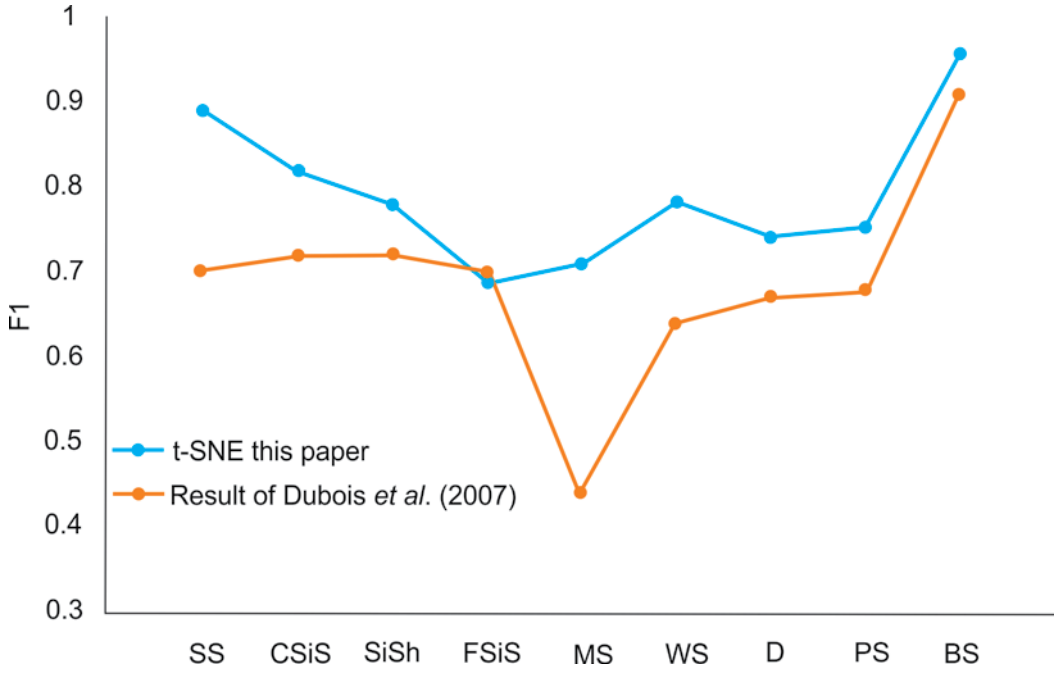

Figure 6 Evaluation metrics for each facies calculated for the classifications with t-SNE+K-NN (in blue) and based on the artificial neural network used by [6] (in orange).

that the blue line is always above the orange line, exhibiting the lowers differences in the SiSH, PS and BS facies.

We also present a visual inspection analysis in Figure 7 and Figure 8 corresponding to the classification of samples belonging to the NOLAN and SHRIMPLIN wells, respectively, organized with respect to the Well depth. In these figures, from left to right, it can be seen the behavior of the measures of the four physical properties as the well's depth increases. The last three columns of the figures show in colored horizontal lines the real facies

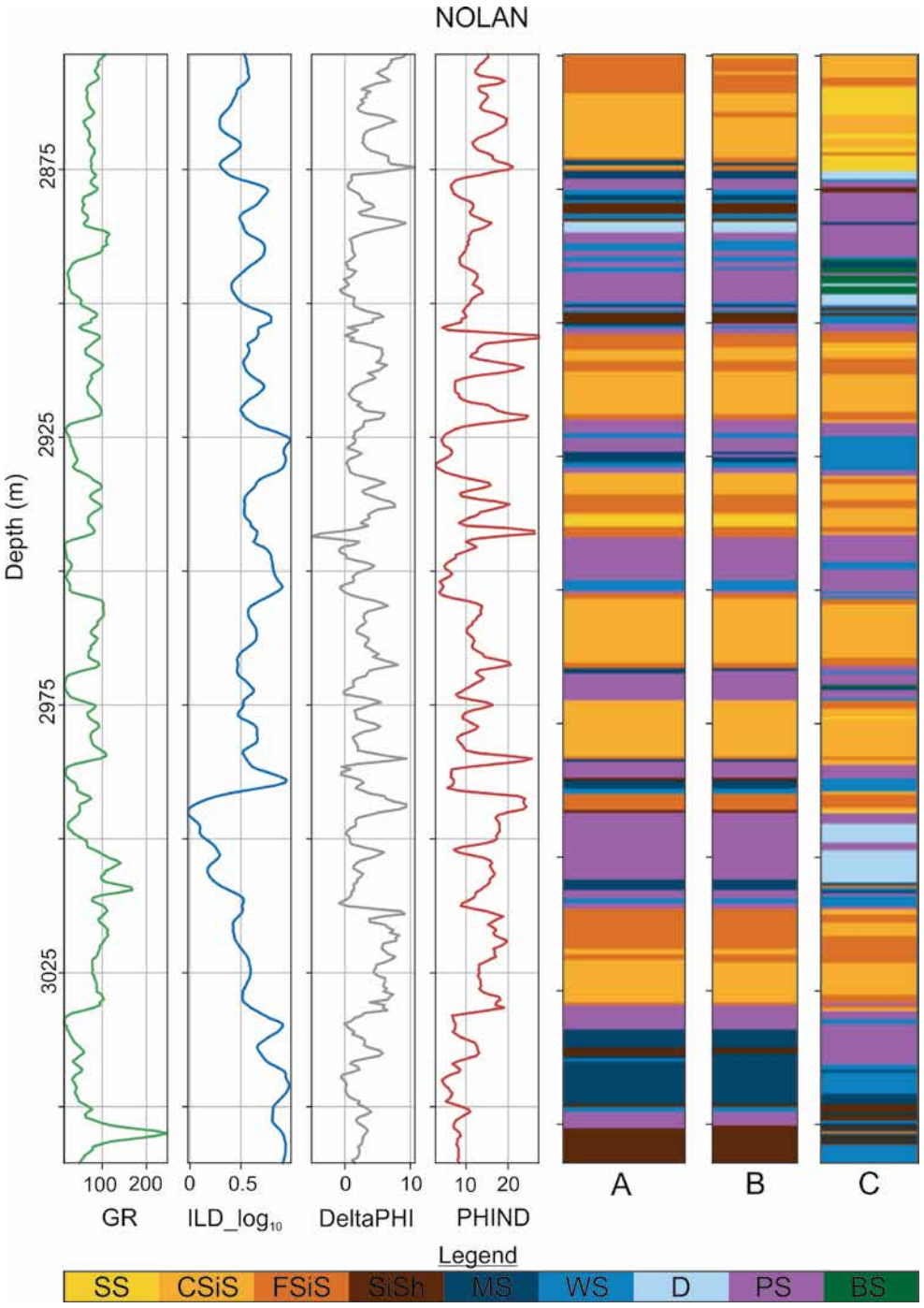

Figure 7 A. Comparison between real facies; B. Facies predicted by t-SNE+K-NN; C. Facies predicted with the Dubois ***et al***. (2007) ANN in the NOLAN. There are also illustrated the gamma ray curves (GR), resistivity on a logarithmic scale on a base 10 basis (ILD_log_{10}), the difference between porosity and neutron density (DeltaPHI) and the average porosity and the neutron density (PHIND).

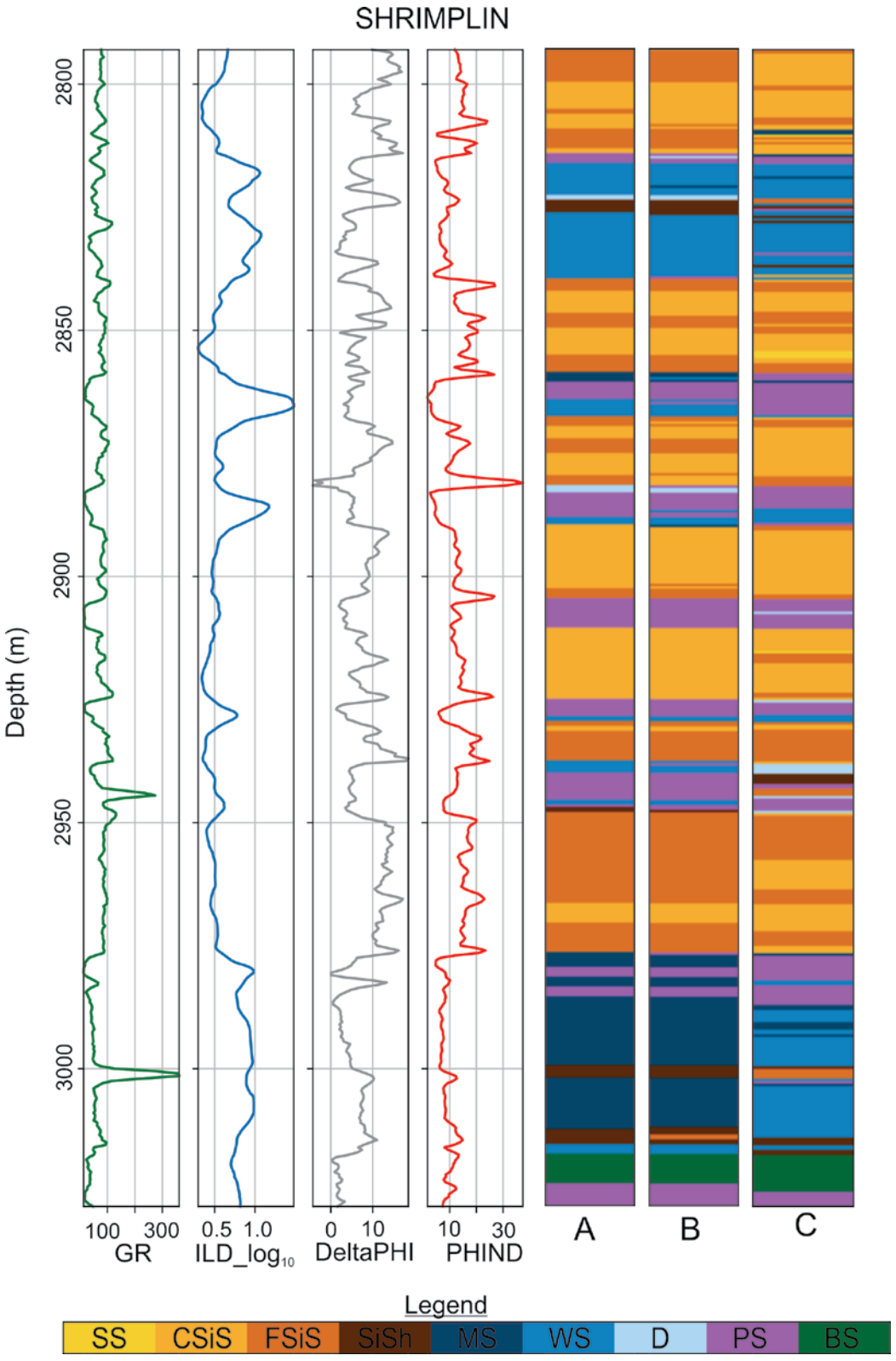

Figure 8 A. Comparison between real facies; B. Facies predicted by t-SNE+K-NN; C. Facies predicted with the Dubois ***et al.*** (2007) ANN in the SHRIMPLIN well. There are also illustrated the gamma ray curves (GR), resistivity on a logarithmic scale on a base 10 basis (ILD_log_{10}), the difference between porosity and neutron density (DeltaPHI) and the average porosity and the neutron density (PHIND).

classes, the predictions made by the t-SNE+KNN, and the ANN approaches, respectively. Examining the behavior of the wells measures variables respecting the real facies classes, it can be noted that abrupt fluctuations in whatever of these variables are usually correlated to changes in the facies classes. Then, this analysis supports the methodology proposed, which considers a sampling window of measures as input features for learning the t-SNE mapping, instead of the methodology followed by Dubois *et al*. (2007).

By analyzing the predictions of the methods presented in the last two columns of Figure 7 and Figure 8, it can be observed that the t-SNE+K-NN graph is almost indistinguishable than the corresponding to the real facies. Contrary, the graph corresponding to the ANN predictions presents remarkable discrepancies with regards to the reference. These results show consistency in our proposed method in getting better performance than the baseline. It is important to emphasize that for the t-SNE+K-NN experiments, we just selected 50% of the training samples for making the predictions. It demonstrates the capability of the t-SNE+K-NN for capturing complex patterns so that samples belongings to the same class are mapped closer in a defined Euclidian feature space.

Finally, Table 5 presents the t-SNE+K-NN confusion matrix organized according to the genesis of the rock groups. Results demonstrated that the t-SNE+K-NN was able to capture the patterns associated with the genesis of rock. It can be discerned that most of the misclassifications between classes disappear. For group 1, 98:5% of the samples were correctly classified, while for group 3 it was 93:9%. Note that, group 2 and group 4 represent the SiSn and BS facies, respectively, such that the performance was the same, i.e., 73:7%, and 3:7%, respectively.

Two elements can be approached to explain differences among lithofacies in terms of predictive capacity. The first is related to an unbalance in the lithofacies of the test set and the training set, which is expected, as they are piles of lithofacies that depend on several geological conditions. Facies with a larger number of label samples offer a greater predictive capacity. The second element is related to data variability in each lithofacies. In all facies, there is log data overlapping, nevertheless, among carbonate facies, data overlapping is higher, with little or almost no difference among average properties in those facies. In non-marine facies, however, average properties in logs present at least some differences among lithofacies.

The difference in performance between the proposed method and the baseline relies on the capability of t-SNE for projecting into a close Euclidean distance the samples that follow a similar pattern. Due to this process is performed unsupervised, the algorithm is not affected by possible errors of labeling, providing a better capacity of generalization, and requiring less labeled samples when using with the K-NN. However, the artificial neural network, based on multi-layer perceptions (MLP), requires a great amount of label data to train the prediction model. Dubois *et al*. (2007) used few indicators in their MLP, which made this network incapable of capturing complex or non-linear relationships within datasets. Expanding the amount of MLP indicators would require even more label samples, and it would run the risk of overfitting. Another important point to make is that Dubois *et al*. (2007)'s MLP could be affected by incorrectly labelled samples, while the model based on t-SNE + K-NN does not offer that risk, as it is an unsupervised model.

Apart from the methodology presented in this study, there are other methodologies based on machine learning targeting prediction of lithofacies, such as Support Vector Machine (SVM), Random Forest (RF), Bayesian Network (BN) and even Artificial Neural Networks (ANN) in different settings (Dubois *et al*., 2007; Al-Anazi & Gates, 2010; Salehi & Bizhan, 2014; Sebtosheikh & Salehi, 2015;

Table 5 Confusion matrix with separation between the proposed groups based on petrogenetic criteria. The hits in group 1 are represented in blue, for group 2 in gray, for group 3 in yellow and for group 4 in green.

		Predict									
		SS	CSiS	FSiS	SiSh	MS	WS	D	PS	BS	Total
True	SS	67	4		1						72
	CSiS	7	202	39	1	1		1	1		252
	FSiS	3	36	158	1				2		200
	SiSh		1	1	42	2	5	1	5		57
	MS	1		1	4	50	3	3	11		73
	WS			2	8	4	113	1	21		149
	D			1			2	26	6		35
	PS			1	7	9	17	2	129	1	166
	BS					1	1	1	1	59	63

Bhattacharya & Carr, 2016; Mishra & Datta-Gupta, 2017; Bhattacharya & Mishra, 2018). Notwithstanding, except for Bayesian Network, other methodologies, although robust, fail to overcome the conception of the correlation coefficient as a measure of accuracy, which makes it hard to classify sets of complex and non-linear data in 2-D space.

Despite the satisfactory results obtained with the t-SNE+KNN approach, the classification of facies using data from wells is still a complicated problem, with high ambiguity, and solutions that can be unrealistic. This is because of the properties measured in well present uncertainties on three levels: for each facies, there exists a wide range of responses for the properties measured, each tool presents different intervals of measuring, and the conditions of the well change.

6 Conclusions

The results presented in this study demonstrate the efficacy of the combination of an algorithm of dimensional reduction like t-SNE with an unsupervised classifier (K-NN) for the prediction of lithofacies in diversified reservoirs, as in the cases of the Hugoton and Panoma fields, which represent repeated vertical successions of lithofacies of cyclic nature, with a rising pattern, and resulting from quick floatation at relative sea level.

The results demonstrate the ability of the proposed method in comparison with the baseline; the t-SNE+K-NN consistently outperformed the baseline up to 10 for all evaluated performance metrics. Most importantly, these results were obtained using 50 of the training labeled samples employed by the ANN approach. The small amount of labelled data (identified lithofacies) necessary to train the model based on the combination of t-SNE + K-NN makes this methodology ideal for fields with a small amount of well cores. It is important to say that, due to the lack of public-domain data from other types of reservoirs, this methodology could not be tested for different types of reservoir, such as reservoirs that are exclusively siliciclastic, leaving that stage for future studies.

The success of the method is due to the t-SNE+K-NN ability to deal with high dimensional and nonlinearly correlated data. It can find patterns embedded in the dataset that are not found by traditional facies prediction methods. Considering the nature of the data and the nature of the problem to be solved, t-SNE+K-NN proved to be a robust tool for the prediction of facies in wells without description of cores.

7 Acknowledgements

The authors would like to acknowledge Petrobras, for research financial support. In addition, we are thankful to Intel Semiconductors Brazil for recognizing ICA Laboratory at PUC-Rio University as Intel AI Innovation Center. To PUC-Rio and Intel Corporation, without which this work could not have been accomplished.

8 References

Aibar, S.; González-Blas, C.B.; Moerman, T.; Imrichova, H.; Hulselmans, G.; Rambow, F.; Marine, J.C.; Geurts, P.; Aerts, J. & Oord, J.V.D. 2017. Scenic: single cell regulatory network inference and clustering. *Nature methods*, *14*(11): 1083–1086.

Al-Anazi, A. & Gates, I.D. 2010. On the capability of support vector machines to classify lithology from well logs. *Natural Resources Research, 19*(2): 125-139.

Albuquerque, C.F.; Soares, J.A. & Bettini, C. 2005. The use of well logs in logfacies modeling–example in the Namorado field, Campos Basin, Brazil. *In*: 9th INTERNATIONAL CONGRESS OF THE BRAZILIAN GEOPHYSICAL SOCIETY & EXPOGEF, Salvador, 2005. Society of Exploration Geophysicists and Brazilian Geophysical Society, p. 1157-1161.

Bhattacharya, S. & Carr, T.R. 2016. Integrated petrofacies characterization and interpretation of depositional environment of the Bakken Shale in the Williston basin, North America. *Petrophysics*, *57*(2): 96-111.

Bhattacharya, S. & Mishra, S. 2018. Applications of machine learning for facies and fracture prediction using Bayesian Network Theory and Random Forest: Case studies from the Appalachian basin, USA. *Journal of Petroleum Science and Engineering*, *160*: 1005-1017.

Burke, J.A.; Campbell Jr, R.L. & Schmidt, A.W. 1969. The litho porosity cross plot: A new concept for determining porosity and lithology from logging methods. *In*: SPWLA 10th ANNUAL LOGGING SYMPOSIUM, 1969, Society of Petrophysicists and Well-Log Analysts.

Busch, J.M.; Fortney, W.G. & Berry, L.N. 1987. Determinação da litologia a partir de perfis de poços por análise estatística. *Avaliação de Formação de SPE*, *2*(4): 412–418.

Cunha, E.S.; Oliveira, K.A. & Gomes, H.M. 2003. Investigação do treinamento de uma rede neural para o reconhecimento de litofácies combinando dados de testemunhos e perfis de poços de petróleo. *In*: CONGRESSO BRASILEIRO DE P&D EM PETRÓLEO & GÁS, 2, 2003, p. 1–6.

Delfiner, P.; Peyret, O. & Serra, O. 1987. Automatic determination of lithology from well logs. *SPE Formation Evaluation*, *2*(03): 303–310.

Dubois, M.; Bohling, G.; Byrnes, A. & Seals, S. 2003. Extracting lithofacies from digital well logs using artificial intelligence, Panoma (council grove) field, Hugoton embayment, Southwest Kansas. *In:* PROCEEDINGS, MID-CONTINENT SECTION AMERICAN ASSOCIATION OF PETROLEUM GEOLOGISTS MEETING, 2003, Tulsa, p. 30.

Dubois, M.K.; Byrnes, A.P.; Bohling, G.C. & Doveton, J.H. 2006. Multiscale geologic and petrophysical modeling of the giant Hugoton gas field (Permian), Kansas and Oklahoma, USA. *In:* HARRIS, P.M. & WEBER, L.J. (eds.). *Giant Hydrocarbon Reservoirs of the World, from Rocks to Reservoir Characterization and Modeling.* American Association of Petroleum Geologists Memoir, 88, p. 307-353.

Dubois, M.K.; Bohling, G.C. & Chakrabarti, S. 2007. Comparison of four approaches to a rock facies classification problem. *Computers & Geosciences*, *33*(5): 599-617.

Heyer, J.F. 1999. Reservoir characterization of the Council Grove Group, Texas County, Oklahoma, *In:* MERRIAM, D.F. (ed.)., AAPG MIDCONTINENT SECTION MEETING TRANSACTIONS, Geosciences for the 21st Century, p. 71-82.

Hsieh, B.; Lewis, C. & Lin, Z. 2005. Lithology identification of aquifers from geophysical well logs and fuzzy logic analysis: Shui-lin area, Taiwan. *Computers & Geosciences*, *31*(3): 263-275.

Kluth, C.F. 1986. Plate tectonics of the ancestral Rocky Mountains, *In*: PETERSON, J.A. (ed.). *Paleotectonics and Sedimentation of the Rocky Mountains, United States*. AAPG Memoir, 41, p. 353–369.

Linderman, G.C. & Steinerberger, S. 2019. Clustering with t-SNE, provably. *SIAM Journal on Mathematics of Data Science*, *1*(2): 313-332.

Maeten, L.V.D. & Hinton, G. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*: 2579–2605.

Mishra, S. & Datta-Gupta, A. 2017. Applied Statistical Modeling and Data Analytics. *Elsevier*. 237p.

Olson, T.M.J.A.; Babcock, K.V.K.; Prasad, S.D.; Boughton, P.D.; Wagner, M.K.; Franklin, M.H. & Thompson, K.A. 1997. Reservoir characterization of the giant Hugoton gas field, Kansas. *AAPG Bulletin*, *81*: 1785-1803.

Perry, W.J. 1989. Tectonic evolution of the Anadarko basin region, Oklahoma. U.S. *Geological Survey Bulletin*, *1866*: 1-16.

Piegl, L.A. & Tiller, W. 2002. Algorithm for finding all k nearest neighbors. *Computer-Aided Design*, *34*(2): 167-172.

Rogers, S.J.; Fang, J.H.; Karr, C.L. & Stanley, D.A. 1992. Determination of lithology from well logs using a neural network. *AAPG Bulletin*, *76*(5): 731-739.

Rosa, H.; Suslick S.B.; Vidal, A.C. & Sakai, G.K. 2008. Electrofacies characterization using multivariate statistical tools. *Revista Escola de Minas*, *61*(4): 415–422.

Salehi, M. & Bizhan, H. 2014. Automatic identification of formation lithology from well log data: a machine learning approach. *Journal of Petroleum Science Research*, *3*(2):73-82.

Sebtosheikh, M.A. & Salehi, A. 2015. Lithology prediction by support vector classifiers using inverted seismic attributes data and petrophysical logs as a new approach and investigation of training data set size effect on its performance. *Journal of Petroleum Science and Engineering*, *134*: 143-149.

Shaham, U. & Steinerberger, S. 2017. Stochastic neighbor embedding separates well-separated clusters. ArXiv preprint arXiv:1702.02670.

Received: 28 May 2020

Accepted: 21 October 2020

How to cite:

Potratz, G.L.; Canchumuni, S.W.A.; Castro, J.D.B.; Potratz, J. & Pacheco, M.C.C. 2021. Automatic Lithofacies Classification with t-SNE and K-Nearest Neighbors Algorithm. *Anuário do Instituto de Geociências, 44*: 35024. DOI 10.11137/1982-3908_2021_44_35024