# Generation of Well Logs using Machine Learning algorithms in different fields of the Cuenca Del Valle Medio Del Magdalena in Colombia

Jorge Torres Arboleda*, Emanuel Chaverra Zuleta, Luis Fernando Duque Gomez, Franco Bertaiola, and Andrés Mauricio Muñoz García.
Grupo de Geofísica y Ciencias de la Computación (GGC3), Instituto Tecnológico Metropolitano de Medellín (ITM)

This paper was prepared for presentation during the 17th International Congress of the Brazilian Geophysical Society held in Rio de Janeiro, Brazil, 16-19 August 2021.

## Abstract

The application of machine learning in recent decades has generated a digital transformation in different areas of knowledge from medicine to physics, including the social sciences. In geosciences it has been used more and more in recent years as it provides new methodologies based on high-performance computing resources and cloud computing, lowering costs and simplifying prospecting processes. This research aims to explore the application of these new technologies in the reconstruction of well logs.

In the present work, four wells located in the Middle Magdalena Valley basin in Colombia are specifically used, wells W1, W2, W3 belonging to the same field and well L located in a neighboring field at approximately 16 km.

The methods of Random Forest, Gradient Boosting, and Artificial Neural Networks were used to train models based on the logs of the first three wells that comprise Gamma Ray (GR), Spontaneous Potential (SP), Sonic (DT), and Density (RHO) to later reconstruct the Compressional wave delay time (DTCO) and Shear wave delay time (DTSM) in the neighboring field. We propose a generalized computational methodology to obtain reliable models from known logs to be applied to distant wells located in fields with different structural conditions, but with lithological similarities. For this, the models trained with data from W1, W2, and W3 on L were applied with control logs that allowed a validation of the method and confirmed the possibility of generating the missing logs to improve the well modeling.

## Introduction

Since the beginning of Colombia's oil exploration activity, hundreds of wells have been drilled, in some of which logs have been taken only in the specific depth range of a geological target, but due to the depletion of reserves, there is a need to reconstructing the logs in the missing areas or registering those that were not taken, which implies an economic investment so high that it would make many projects unviable.

An attractive proposition in the era of digital transformation and the fourth industrial revolution is the use of computational resources such as artificial intelligence, specifically Machine Learning (ML) to support all areas of knowledge and scientific and industrial processes.

Currently, ML is permeated by almost all human activities and is used in a wide range of situations, such as in social networks to generate recommendations for products, services or contacts to its users, based on their tastes and intuited preferences. from the data analysis of their previous behavior in the network (Dehuri et al., 2014; Kalyanam, 2017; Nurek & Michalski, 2020). Image recognition is also used using deep learning algorithms to find and label faces, this same type of algorithm can be used in medicine to detect abnormalities in X-ray images, CT scans and other medical images, to automatically diagnose diseases (Cleophas & Zwinderman, 2013; Sagar, 2019).

Geosciences are not immune to the influence of these computational methods in the development of some of their activities, for example, in hydrocarbon prospecting, a discipline in which important information is generated through different techniques such as reflection seismic and well logs to classify lithological structures or formations of geological interest, ML algorithms have been implemented to solve different problems and meet a particular objective, in the case of waveform recognition and first arrivals, analysis of well logs (Huang et al., 1996) or seismic tomography (Araya-Polo et al., 2018), estimation of rock parameters at a seismic scale using well logs, seismic data and Artificial Neural Networks (ANN), among others (Helle et al., 2001; Iturrarán-Viveros, U. & Muñoz-García AM, 2018; Iturrarán-Viveros, U., Muñoz-García, AM, & Parra, J. O, 2018; Iturrarán-Viveros, 2012; Iturrarán-Viveros & Parra,2014).

Computational ML-based methods for well log reconstruction face an additional challenge when insufficient proprietary data are available to develop reliable ML models. Therefore, in this work, we apply a new computational methodology that uses data from wells from different areas, but with similar lithologies to allow the evaluation of different Machine Learning methods and obtain simplified tools as a resource in geosciences that are implemented with an established logical sequence.

The development and implementation of the algorithms were carried out on four wells located in the middle Magdalena valley basin in Colombia, where one of them (L) does not have part of its DTSM (S-wave dipole) and DTCO (P-wave dipole) logs. In order to develop an algorithm capable of reconstructing these logs from data from other wells (W1, W2 and W3) located at distances of approximately 14, 16 and 17 kilometers respectively from L in a different oil field (Figure 1), the models were elaborated from Random Forest (RF), Gradient Boosting (GB) methods and artificial neural networks, these were previously validated on each W well to later be applied on L with a systematic methodology that allows evaluating in an agile way each model and build systematized algorithms.

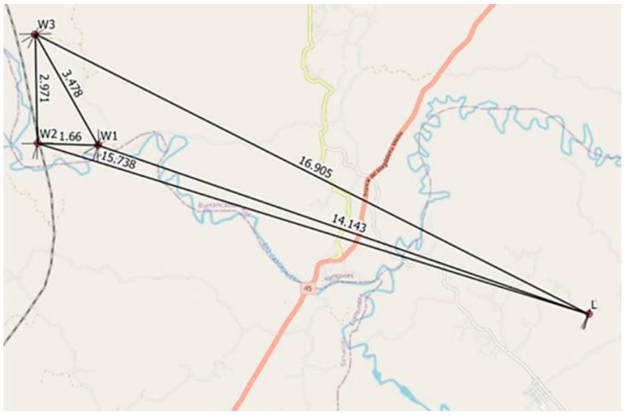


***Figure 1** – Study area location map (distance in meters)*

## Method

Our methodology was based on data from existing rock parameters resulting from the analysis of the drilling cores and comprised three phases (Figure 2). In the first phase, data preprocessing was carried out, this allowed to build an optimal and efficient database for the training of machine learning algorithms. The second phase consisted of comparing different ML methods based on statistical metrics through the development of algorithms applied on test data to obtain a competent machine learning model for each parameter, finally, in the third phase the algorithm obtained was implemented to generate the objective parameter and evaluate its competence by comparing with the real data and previous metrics, giving rise to a computational methodology for the reconstruction of well logs. The different methods were applied to well L and it was determined which of these provides relevant information on the logs, which allows the elaboration of modules for specific cases based on the specific parameter that is required to be predicted and the structural complexity of the study area.

The data available from the W wells includes real P-wave (DT) sonic logs, spontaneous potential (SP), density (RHO), deep induction resistivity (ILD), real and synthetic dipole S-wave logs (DTSM), p-wave dipole (DTCO), gamma rays (GR) in addition to the depth to which each record corresponds.

The target well L has complete logs of DT, SP and RHO logs, but lacks some logs of DTCO and DTSM, the methodology was applied for each of these incomplete data to obtain a reliable model for its reconstruction from known data.

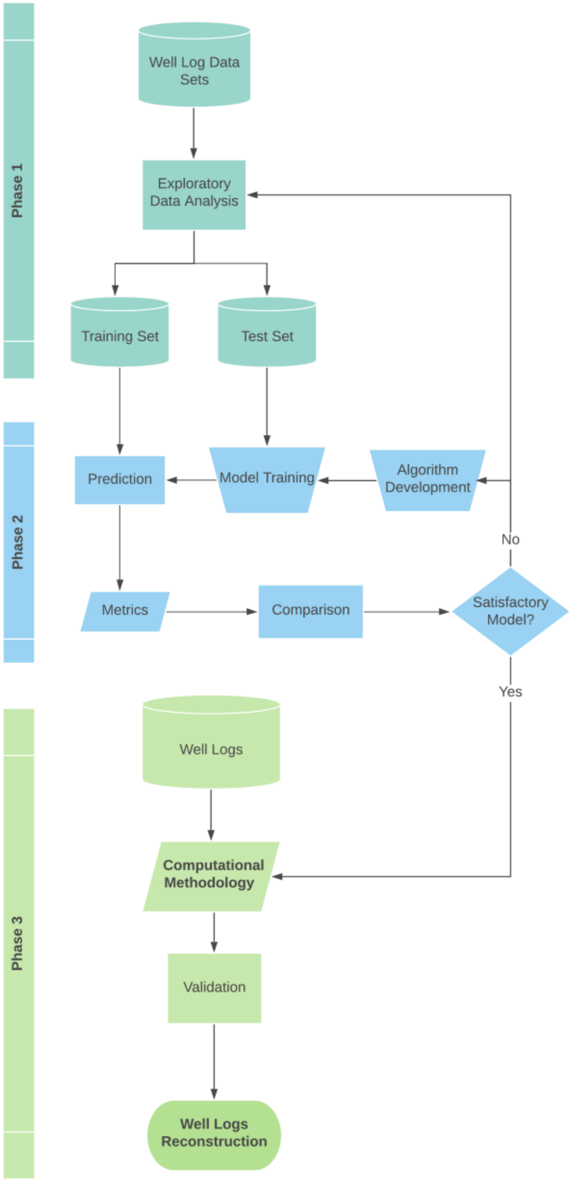


***Figure 2** – Development methodology*

Each methodological phase is described below:

**Phase 1:**

In the preprocessing, an exploratory analysis of the data and the respective manipulation was carried out to homogenize the structure (Table 1) so that it agrees with the input parameters as a function of the parameter to be reconstructed.

**Table 1** – Example data structure.

| DEPTH (ft) | DT (us/ft) | SP (mv) | RHO (g/cm3) | GR | DTCO (us/ft) |
|---|---|---|---|---|---|
| 53 | 112.55 | -89 | 2.3279 | 44.15 | 112.55 |
| 53.5 | 114.88 | -89.25 | 2.3279 | 43.94 | 114.88 |
| 54 | 99.62 | -89.63 | 2.3279 | 42.93 | 99.62 |

For the DTSM and DTCO reconstruction cases, it was determined that the input parameters would be DT, SP, RHO and GR. In the case of GR, they would be DT, SP and RHO.

The other logs were discarded from the database, including depth, given the purpose of the models requires that this parameter does not generate a bias in the reconstructions.

3 data sets were created::

1. Data from wells W
2. Data from well L
3. Mixed data from sets 1 and 2 (WL)

Each set was subjected to the following phases of the methodology.

Once the appropriate configuration is had where the input parameters and the output parameter are determined, a random division of the data is made into 80% of data for model training and 20% for evaluation.

**Phase 2:**

An algorithm was developed for each ML model to be used (Random Forest, Gradient Boosting and Artificial Neural Networks), this consisted in determining the appropriate hyperparameters for each case.

The models were trained with the previously separated data and internally built a series of relationships that allow you to learn to determine the objective from the input parameter data.

Then, the predictions are made with the test data, which were not used in the training, and retain their objective parameter, this allows obtaining error statistics between the reconstructed data and the real data, based on the evaluation of the metrics, for this, the necessary modifications are made to the algorithm and the input data, seeking to obtain a competent model.

**Phase 3:**

The model becomes part of the computational methodology which only requires defining the input data and the objective parameter. The methodology and the model it contains are evaluated taking all the data from L that contain real target logs to generate predictions and make comparisons from which new metrics are obtained, not only is it sought to reduce the error between prediction and real data, but also compare the metrics with those obtained in the test predictions to demonstrate the reliability of the model and guarantee that the error statistics are constant with new data, in this way it can be stated that the reconstructed logs of L that do not count the objective parameter to be evaluated by the error between prediction and real data, they have an error similar to the error established with the model.

The metrics used were the mean square logarithmic error (MSLE), the mean absolute error (MAE) and the coefficient of determination ($R^2$).

**Results**

The proposed methodology was effective to develop the models and algorithms to carry out the exploration, it is important to highlight that the dataset was modified during the process, eliminating data that generate biases in the models.

The ANN model trained with the W wells is not adequate to predict DTSM or DTCO in L, since the metrics diverge too much (Table 2), which indicates that the model does not have enough data to achieve the required generalization and to be able to predict a well from completely external data, this is corroborated by observing the results of the ANN model trained with data from L (Table 3), which as expected has an outstanding result when predicting data from the same well with which it was trained (Figure 3(a)).

**Table 2** – ANN model trained with W dataset

| ANN W | DTSM | | DTCO | |
|---|---|---|---|---|
| Métricas | Modelo | L | Modelo | L |
| MSLE | 0.011359 | 0.637678 | 0.004658 | 0.5806057 |
| MAE | 13.265303 | 254.288788 | 4.143526 | 211.9466 |
| $R^2$ | 0.630605 | -26.636583 | 0.553510 | -374368.1 |

**Table 3** – ANN model trained with L dataset

| ANN L | DTSM | | DTCO | |
|---|---|---|---|---|
| Métricas | Modelo | L | Modelo | L |
| MSLE | 0.010127 | 0.011026 | 0.003530 | 0.003443 |
| MAE | 16.014362 | 16.556129 | 3.277894 | 3.229229 |
| $R^2$ | 0.903947 | 0.900856 | 0.924684 | 0.929315 |

By mixing the data of W and L, the neural networks achieve greater generalization, and the results are close to those of the model trained with L, losing precision, but maintaining a good relationship between their metrics (Table 4, Figure 3(c)).

**Table 4** – ANN model trained with WL dataset

| ANN WL | DTSM | | DTCO | |
|---|---|---|---|---|
| **Métricas** | **Modelo** | **L** | **Modelo** | **L** |
| **MSLE** | 0.011841 | 0.010917 | 0.003988 | 0.003391 |
| **MAE** | 15.443144 | 16.314772 | 3.689694 | 3.222662 |
| **R²** | 0.881985 | 0.903446 | 0.896464 | 0.931069 |

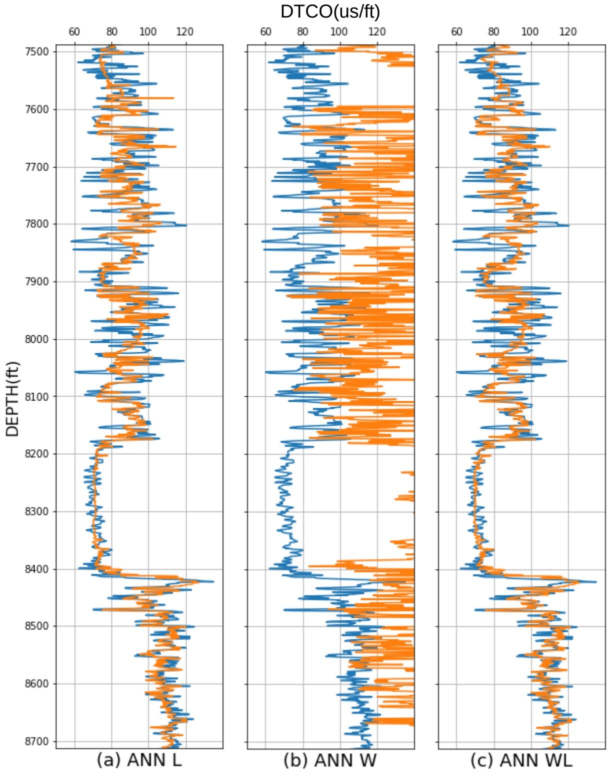


***Figure 3*** *– DTCO prediction from ANN models, real logs in blue, synthetic logs in orange, (a) Model trained with L, (b) Model trained with data from W and L, (c) Model trained with data from W*

Both Gradient Boosting and Random Forest trained with the W wells cannot predict high DTCO and DTSM values, This is due to the fact that the training data has values lower than L values (Figure 4), however, below these values a higher performance was recorded than neural networks (Figure 5(b)), considering that the model was not trained with any of the well logs to predict, it is an interesting case that indicates the possibility of using computational methodologies to reconstruct Well logs from different wells provided that lithological similarities can be guaranteed, further studies are required to fully determine the necessary conditions in which these models operate adequately in this methodology.

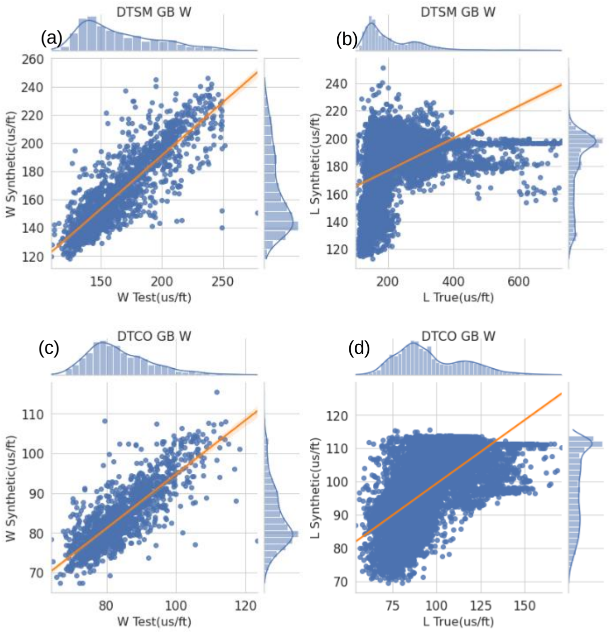


***Figure 4*** *– Scatter Plots, Comparison of w test data and L data, (a) DTSM GB W test, (b) DTSM GB W applied to L, (c) DTCO GB W test, (d) DTCO GB W applied to L*

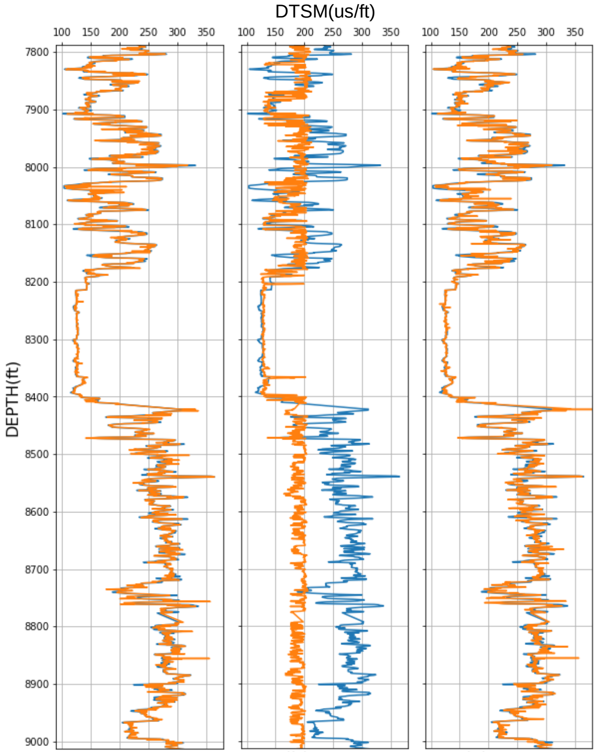
**Table 5** – GB model trained with W dataset

| GB W | DTSM | | DTCO | |
|---|---|---|---|---|
| **Métricas** | **Modelo** | **L** | **Modelo** | **L** |
| **MSLE** | 0.008378 | 0.111748 | 0.003506 | 0.021508 |
| **MAE** | 10.9657 | 55.68066 | 3.599727 | 11.602841 |
| **R²** | 0.732031 | 0.006333 | 0.671646 | 0.403711 |

**Table 6** – RF model trained with W dataset

| RF W | DTSM | | DTCO | |
|---|---|---|---|---|
| **Métricas** | **Modelo** | **L** | **Modelo** | **L** |
| **MSLE** | 0.009326 | 0.091782 | 0.003644 | 0.019980 |
| **MAE** | 11.498577 | 52.10701 | 3.56904 | 11.162438 |
| **R²** | 0.705852 | 0.129556 | 0.649939 | 0.401343 |

As for the GB and RF models trained with L's own data, an almost perfect precision is appreciated, with outstanding metrics (Table 7, Table 8), especially in GB, which had the best performance (Figure 4(a)), even reducing the error when applied to the data. full of L, this is because there is both the test data of the model and the training data, although the risk of overfitting is present, the error metrics of the model obtained from the test data are kept low, this makes the model obtained useful for reconstructing logs in well L but not general enough to be applied in other wells unless the lithology is very similar, as is the case with models trained only with W to predict L.



(a) GB L (b) GB W (c) GB WL

***Figure 4** – DTSM prediction from GB models, real records in blue, synthetic records in orange, (a) Model trained with L, (b) Model trained with data from W, (c) Model trained with data from W and L*

**Table 7** – GB model trained with L dataset

| GB L | DTSM | | DTCO | |
|---|---|---|---|---|
| **Métricas** | **Modelo** | **L** | **Modelo** | **L** |
| **MSLE** | 0.006184 | 0.001423 | 0.002419 | 0.000552 |
| **MAE** | 11.424688 | 3.889991 | 2.564090 | 0.932854 |
| **R²** | 0.943789 | 0.988537 | 0.951145 | 0.988696 |

**Table 8** – RF model trained with L dataset

| RF L | DTSM | | DTCO | |
|---|---|---|---|---|
| **Métricas** | **Modelo** | **L** | **Modelo** | **L** |
| **MSLE** | 0.005493 | 0.001099 | 0.002082 | 0.000437 |
| **MAE** | 10.640422 | 2.133635 | 2.373364 | 0.665016 |
| **R²** | 0.959602 | 0.991582 | 0.956601 | 0.990879 |

All the models trained with the combined data show an intermediate performance between the models trained only with L or W, being higher than the W models and lower than the L models, behaving according to expectations, better with the known data and worse with the data from different wells, showing the importance not only of the amount of data but also of their nature, it should be noted that the most accurate and precise model is not always the ideal since it will depend on the conditions in which the model will be tested, in our work we intend to explore the possibility of obtaining a generalized model that is good enough to reconstruct well logs using data from neighboring wells for which a precise model is not viable but limited to its own logs, in this order of ideas the most favorable results in general correspond to the WL data (combination of data from wells W and L) and they open the possibility of establishing a methodology that consists of accumulating data from suitable wells according to lithology to perform reconstructions that only require adding a small sample of the incomplete well and thus overcome the challenge posed by wells with too few logs to provide sufficient data in the development of their own models.

**Table 9** – GB model trained with WL dataset

| GB WL | DTSM | | DTCO | |
|---|---|---|---|---|
| **Métricas** | **Modelo** | **L** | **Modelo** | **L** |
| **MSLE** | 0.007544 | 0.001617 | 0.002781 | 0.000924 |
| **MAE** | 11.604986 | 4.740306 | 2.99111 | 1.57143 |
| **R²** | 0.940969 | 0.989035 | 0.928758 | 0.980929 |

**Table 10** – RF model trained with WL dataset

| RF WL | DTSM | | DTCO | |
|---|---|---|---|---|
| **Métricas** | **Modelo** | **L** | **Modelo** | **L** |
| **MSLE** | 0.006666 | 0.0010696 | 0.002623 | 0.000407 |
| **MAE** | 10.579103 | 2.0408520 | 2.841404 | 0.669509 |
| **R²** | 0.951921 | 0.9929056 | 0.930047 | 0.991817 |

## Conclusions

The use of computational resources such as artificial intelligence, specifically Machine Learning (ML) is an attractive proposal in the era of digital transformation and the fourth industrial revolution to support all areas of knowledge, scientific and industrial processes.

The relevance of the volume of data to adequately train a ML model is evidenced, as well as its nature.

The results indicate that artificial neural networks may be biased due to the nature of the training data, an established model can be well adapted to be used in a specific situation such as the reconstruction of well logs from own data, but requires a large amount of data and creates problems when dealing with wells with different lithologies, which is also supplemented by adding data from these wells to the training dataset.

Despite the popularity of neural networks in general use, other ML methods showed greater generalizability and therefore more promising results when trained with data external to the study well, giving consistent results where the lithologies were similar. More studies are necessary to establish the conditions where this is useful, but it is evident that it responds to geological characteristics that can be determined to develop the appropriate algorithm for each well from known neighboring wells.

In the exploration and application of the methodology, Gradient Boosting regression was able to adequately predict specific ranges from well data completely external to the data of its training, indicating that the model is viable to elaborate a reconstruction module oriented to specific lithologies that indicate the presence or absence of particular properties, the potential that the exploration of this method can have can be appreciated by expanding the training database in a controlled manner.

## Acknowledgments

We thank the Instituto Tecnológico Metropolitano de Medellín (ITM), project code P20248, project code P21111, Ecopetrol S.A. and Colciencias for the information and data provided under the 0266-2013 Colciencias-Ecopetrol program..

## References

Araya-Polo, M., Jennings, J., Adler, A., & Dahlke, T. (2018). Deep-learning tomography. The Leading Edge, 37(1), 58-66. https://doi.org/10.1190/tle37010058.1

Cleophas, T. J., & Zwinderman, A. H. (2013). Machine learning in medicine - a Complete Overview. Springer. https://doi.org/10.1007/978-94-007-5824-7.

Dehuri, S., De, S., & Wang, G. (2014). Machine Learning for Social Network Analysis: A Systematic Literature Review. The IUP Journal of Information …, 3(4), 30–51. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2187186.

Helle, H. B., Bhatt, A., & Ursin, B. (2001). Porosity and permeability prediction from wireline logs using artificial neural networks: A North Sea case study. Geophysical Prospecting, 49(4), 431–444. https://doi.org/10.1046/j.1365-2478.2001.00271.x

Huang, Z., Shimeld, J., Williamson, M., & Katsube, J. (1996). Permeability prediction with artificial neural network modeling in the Venture gas field, offshore eastern Canada. Geophysics, 61(2), 422-436. https://doi.org/10.1190/1.1443970

Iturrarán-Viveros, U. (2012). Smooth regression to estimate effective porosity using seismic attributes. Journal of Applied Geophysics, 76, 1–12. https://doi.org/10.1016/j.jappgeo.2011.10.012

Iturrarán-Viveros, U., & Parra, J. O. (2014). Artificial Neural Networks applied to estimate permeability, porosity and intrinsic attenuation using seismic attributes and well-log data. Journal of Applied Geophysics, 107, 45–54. https://doi.org/10.1016/j.jappgeo.2014.05.010

Iturrarán-Viveros, U., & Muñoz-García, A. M. (2018). Porosity and water saturation in sands or shales using Artificial Neural Networks and seismic attributes in a clastic reservoir in Colombia. SEG Global Meeting Abstracts, 1282–1285. https://doi.org/10.1190/igc2018-314.

Iturrarán-Viveros, U., Muñoz-García, A. M., & Parra, J. O. (2018). Petrophysical seismic images obtained with Artificial Neural Networks as prior models for full-waveform inversion: A case study from Colombia. EG Technical Program Expanded Abstracts : 2236-2240.https://doi.org/10.1190/segam2018-2996124.1.

Kalyanam, J. (2017). Machine Learning and Applications on Social Media Data [Disertación para el grado de Doctor de Filosofía, University of California, San Diego]. Recuperada de https://escholarship.org/uc/item/6545w71z.

Nurek, M., & Michalski, R. (2020). Combining machine learning and social network analysis to reveal the organizational structures. Applied Sciences (Switzerland), 10(5). https://doi.org/10.3390/app10051699.

Sagar, A. (2019). Deep Learning for Detecting Pneumonia from X-ray Images. Towards Data Science. Recuperado 14 de junio de 2021 de https://towardsdatascience.com/deep-learning-for-detecting-pneumonia-from-x-ray-images-fc9a3d9fdba8.