minerals

MDPI

Article

Source Rock Evaluation from Rock to Seismic Data: An Integrated Machine-Learning-Based Work Flow and Application in the Brazilian Presalt (Santos Basin)

Maria Anna Abreu de Almeida dos Reis *, Andrea Carvalho Damasceno, Carlos Eduardo Dias Roriz, André Leonardo Korenchendler, Atilas Meneses da Silva, Eric da Silva Praxedes and Vitor Gorni Silva

Petróleo Brasileiro S.A.—PETROBRAS, Av. República do Chile, 65, Rio de Janeiro 20031-170, Brazil; adamasceno@petrobras.com.br (A.C.D.); roriz@petrobras.com.br (C.E.D.R.); andre.korenchendler@petrobras.com.br (A.L.K.); atilasmeneses@petrobras.com.br (A.M.d.S.); epraxedes@petrobras.com.br (E.d.S.P.); vgorne@petrobras.com.br (V.G.S.)

* Correspondence: maria.reis@petrobras.com.br

Abstract: The capacity to predict the occurrence and quality of source rocks in a sedimentary basin is of great economic importance in the evaluation of conventional and non-conventional petroleum resources. Direct laboratory examinations of rock samples are the most accurate way to obtain their geochemical properties. However, rock information is usually sparse, and source rocks are often sampled at positions that may not be representative of the average organic content and quality of oil kitchens. This work proposes a work flow supported by machine learning methods (random forest, DBSCAN, and NGBoost) to automate the source rock characterization process to maximize the use of available data, expand data information, and reduce data analysis time. From the automated quality control of the input data through the extrapolation of laboratory measurements to continuous well logs of geochemical properties, culminating in the 3D estimation of these properties, we generate volumes of total organic carbon (TOC) by applying machine learning techniques. The proposed method provides more accurate predictions, reducing uncertainties in the characterization of source rocks and assisting in exploratory decision making. This methodology was applied in the presalt source rocks from Santos Basin (Brazil) and allowed us to quantify the TOC distribution, improving the interpretation of the main source rock interval top and base based only on seismic amplitude data. The result suggests higher TOC values in the northern and western grabens of the studied area and a higher charge risk in the eastern area.

Keywords: source rock; TOC; machine learning; Santos Basin; presalt

Citation: Reis, M.A.A.d.A.d.; Damasceno, A.C.; Roriz, C.E.D.; Korenchendler, A.L.; Silva, A.M.d.; Praxedes, E.d.S.; Silva, V.G. Source Rock Evaluation from Rock to Seismic Data: An Integrated Machine-Learning-Based Work Flow and Application in the Brazilian Presalt (Santos Basin). *Minerals* **2023**, *13*, 1179. https://doi.org/10.3390/min13091179

Academic Editor: Stavros Kalaitzidis

Received: 14 June 2023
Revised: 4 August 2023
Accepted: 6 August 2023
Published: 8 September 2023

1. Introduction

Risk assessment concerning the effectiveness of a petroleum system's elements and processes plays a major role in petroleum exploration. Source rocks are essential elements for the existence of unconventional resources (shale oil or shale gas) or conventional petroleum accumulations. Our major goal in this work is to assist a more accurate calculation of exploratory risks, particularly the assessment of hydrocarbon charge, by improving the quality and the vertical and lateral resolution of source rock characterization at the basin scale, applying an agile integrated approach using machine learning techniques.

The amount of total organic carbon (TOC) in source rocks depends on the balance between primary productivity, preservation, and mineral dilution controlled by the sedimentation rate. The kerogen quality is related to the type of organic matter deposited in the depositional substrate and the preservation degree, which is mainly regulated by the redox potential in the water column and within sediments. Source rocks under appropriate thermal evolution, as defined by temperature and time, can reach the process of generation and expulsion of petroleum [1].

Traditionally, source rocks are identified and characterized through geochemical analysis of rock and petroleum samples from wells. Rock-Eval pyrolysis is a fast method (approximately 30 min per analysis) and requires a small amount of pulverized rock. It is based on the selective detection and quantification of hydrocarbon and oxygenated compounds released by pyrolysis of organic matter on a predetermined heating schedule [2]. The amounts of hydrocarbons and CO_2 generated are measured as peaks as a function of time and recorded in the form of a pyrogram. Combined with TOC measurements, it is frequently used to measure the quantity, quality, and thermal maturity of organic matter in rock samples [1]. However, as the main objectives of exploration and production wells are reservoir rocks, source rocks samples are usually sparse, and are often sampled at positions that may not be representative of the average organic content or quality of the oil kitchens. To circumvent the limited core sample data, different methods have been proposed to obtain an estimate of the total organic carbon using geophysical well log data (e.g., [3,4]). However, source rock characterization involves the quantification of other geochemical properties, such as the hydrogen index (HI), hydrocarbon potential (S_2), and maturity (e.g., Tmax), which can also be obtained through Rock-Eval pyrolysis. We take all these properties into account in our approach.

Source rocks rich in organic matter tend to have lower density values than non-source rocks with the same mineralogy and burial, in association with higher gamma rays (GR), resistivity (RT), neutron porosity (NPHI), and slowness (DT) (e.g., [3,5,6]). Furthermore, their properties vary in relation to the type and thermal history of the organic matter [3]. Due to the vertical resolution of the logs (from 0.2 to 0.8 m in the cases of the logs used in this work), the source rocks can be identified even when their thickness is below the resolution of the curves used, although the quantification is imprecise.

Although well log data provide information with a relatively good vertical resolution, they are scattered in the basin and provide only local information. Seismic data can provide reliable information to spatially guide the identification of source rocks and the characterization of their geochemical properties, especially their organic content. Seismic inversion is an efficient technique to infer elastic properties of rocks that, in conjunction with geophysical logs and geochemical analysis, allows for an integrated characterization of the potential source rock interval with good vertical and lateral resolution (e.g., [7–10]).

Several authors have applied different methods to characterize the petrophysical and elastic properties of shales rich in organic matter and attempt to predict their occurrence and spatial variation using seismic data (e.g., [11–13]). The conventional seismic approach to source rock characterization is performed by calibrating linear regressions between acoustic impedance (P-impedance, the product of compressional wave velocity and density) and the TOC using well logs and extrapolating these relationships using seismic data to generate volumetric estimations of this property [8].

The presence of organically rich shales tends to reduce the seismic velocities and density and increase the anisotropy in comparison to organic lean shales of similar mineralogy and burial (e.g., [14–16]). Vernik and Nur (1992) [15] considered those changes to be relative to the kerogen content, microstructure, and maturity of the source rock. The physical–chemical interactions with the pore fluids [16] and the pore pressure produced by the conversion of the kerogen into oil [17] can also influence these parameters.

The ratio between compressional (Vp) and shear wave (Vs) velocities (Vp/Vs), which are sensitive to lithology and interstitial fluid, can increase or decrease with the organic content, which can be substantially affected by the variation in the mineral composition of the shales (e.g., [6,18–24]). Because of the lower density and velocity of the organic matter relative to the other minerals present in the rock, the acoustic impedance decreases in a non-linear way with the increase in the organic content [8]. The decrease in acoustic impedance and the increase in anisotropy result in characteristic seismic behavior [10]. The reflections of rich and thick source rocks have high amplitudes when compared to reflections of non-source rocks and are amplified with increased organic content [25], thermal maturation [11],

and porosity [23]. The discrimination of each property that affects the seismic response is a big challenge and is usually achieved with rock physics numerical models (e.g., [26]).

The amplitudes of organically rich shales tend to vary with the offset or angle (AVO/AVA, respectively) in contrast to the host rocks (background) [7,10,11,23,25], which justifies the use of AVO analyses for the characterization of source rocks. Zhu et al. (2011) [19] observed that variation in lithology can significantly influence the AVO response of the source rock, since it is related to the VP/VS ratio. Numerical models of rock physics revealed that the mineralogy of the source rock can influence the class of seismically observed AVO anomaly. Silica-rich source rocks can result in AVO class 3, which occurs when there is a reduction in P-impedance and VP/VS (or Poisson's ratio) between the overlying (enclosing) rock and the source rock. On the other hand, clay-rich source rocks can result in AVO class 4, with reduced P-impedance but with little variation or even an increase in the Poisson ratio between the overlying and source rock.

The verification of the strong correlation between the reduction in the acoustic impedance and the increase in the TOC, combined with the availability of seismic stacks from different ranges of angles or offsets, allows for the quantitative evaluation of the source rock, increasing the reliability in the prediction of its occurrence and distribution [10]. Seismic attributes (any quantity calculated from the seismic data) can also provide relevant information and can highlight amplitude, phase, and frequency changes in seismic data.

Del Monte et al. (2018) [27] compared the signature of source rocks using different methods, including inversion, AVO analysis, and seismic attributes. Despite comparing the results of each methodology, the integration and interpretation of data were performed independently. As each elastic property has a different sensitivity to the petrophysical properties of the rock, the use of multiple properties for TOC prediction instead of just one (as is conventionally done) allows for a reduction in the intrinsic ambiguities between the effects of porosity, pore shape/fracture, and the amount and maturity of organic matter (OM). In this work, we use a broad set of elastic attributes as an input to a machine learning model that relates elastic and geochemical properties. The application of machine learning techniques provides some advantages compared with the aforementioned conventional approaches. For example, machine learning enables the review of large volumes of data to discover specific trends and patterns that would not be apparent to humans, and no human intervention is needed, allowing them to make predictions and improve the algorithms on their own. Moreover, machine learning techniques are good at learning non-linear representations from multidimensional data [28]. Thus, machine learning methods make it possible to obtain more accurate predictions of the properties, reducing ambiguities and enabling a better separation of the previously listed effects, which overlap in the elastic responses of the source rock [29]. After comparing numerous methods, random forest [30], DBSCAN [31], and NGBoost [32] were chosen in this work for their superior performances. A possible pitfall in the seismic estimation of geochemical properties is the ambiguity between the properties of the reservoir and the source rock intervals rich in organic matter. The behavior of elastic properties and electrical well logs may be similar in both situations. The use of multiple attributes tends to reduce this ambiguity. Another possible risk in applying this methodology is overfitting the machine learning model. To minimize these effects, we analyzed graphs of the variation of the metrics in the training, test, and validation data and applied a cross-validation process.

The main objective of this work was to develop an integrated work flow from rock samples to seismic data based on the use of machine learning algorithms for source rock evaluation. The strengths of the proposed methodology are the inclusion of automated quality control of the input data and the estimation of the quantity, quality, hydrocarbon potential, and maturity of the organic matter. The use of geochemical data from rock samples, well logs, and seismic data on a machine learning basis allows us to maximize the use of high-quality data, improving estimates of geochemical properties in terms of assertiveness, efficiency, and speed. The geochemical volumes (with adequate vertical and

lateral resolution) support a more accurate calculation of exploratory risks, notably those related to the assessment of uncertainties regarding petroleum charge assessment.

2. Geologic Setting

The eastern Brazilian basins are classified as continental rift basins and are related to the rupture of the Gondwana Supercontinent and, consequently, the opening of the South Atlantic Ocean. They were formed during the Lower Cretaceous, when a thick succession of continental, fluvial and lacustrine sediments, siliciclastic and carbonate, were deposited in salty and freshwater lake environments controlled by the extensional stresses of the rift phase. Locally, intercalation of volcanic rock can be found. After the rift interval, the thermal subsidence phase began, with features of gravitational slip (e.g., [33]).

The Santos Basin, located in the southeastern Brazilian margin, is the biggest offshore Brazilian basin, with an area of more than 350 thousand square kilometers along the coasts of the sates of Rio de Janeiro, São Paulo, Paraná, and Santa Catarina [34] and is limited by the Cabo Frio structural high in the north and by the Pelotas structural high in the south (Figure 1). Recently, the Santos Basin became the largest producer of oil and natural gas in Brazil due to the discovery of large oil fields, such as Tupi and Búzios, after confirmation of the presalt play in 2006 through the drilling of the wildcat 1-BRSA-329D-RJS (Parati).

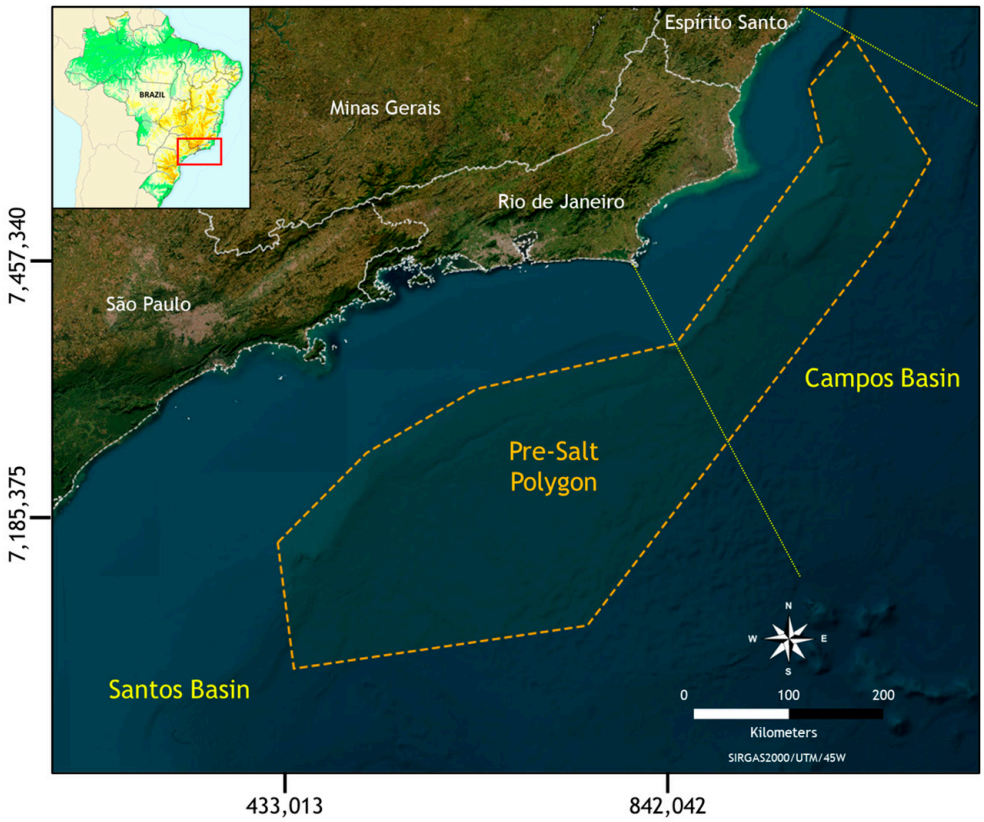

Figure 1. Location of the Santos Basin and the presalt province. The exact location of the research area cannot be shown for confidentiality reasons.

The presalt play is composed of two reservoirs: the main sag reservoir, which was deposited during the early and late Aptian, and the rift reservoir, which was deposited during the Late Barremian and early Aptian (Itapema Formation). The Itapema Formation, corresponding to the Jiquiá Brazilian Local Stage, is characterized by the intercalation of carbonates and black shales [34] and is the main source rock for the hydrocarbon accumulations in the basin. The wells that reached this interval have proven excellent characteristics for petroleum generation, with TOC up to 16% and hydrocarbon potential (S_2) up to 149 mg HC/g rock, possibly related to Ocean Anoxic Event 1a (Freitas et al., 2022 [35] and references therein).

3. Methods

To achieve the intent of this work, an area with a considerable number of wells that acquired data of the Jiquiá source rock (Itapema Formation) and covered by high-quality seismic data was selected. The work flow for source rock characterization is summarized in Figure 2. As presented by Damasceno et al. (2022) [36], the first steps consist of the application of automated quality controls for the measurements of geochemical properties in rock samples and well logs. Next, we fit a machine learning model relating the basic suite of well logs to the measurements of geochemical properties from the automatically validated rock samples. The final step is the fit of a model correlating elastic properties and the TOC well logs predicted in the previous step.

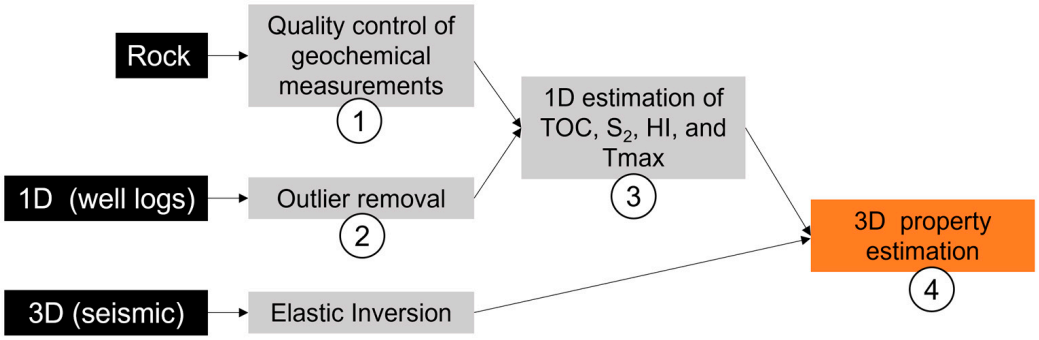

Figure 2. Work flow for automated source rock assessment. The numbers correspond to the machine-learning-based algorithms described in this work.

The methodology used in the application of each step of the proposed flow is described separately for each algorithm as follows.

3.1. Algorithms 1 and 2: Quality Control of the Input Data

The use of poor-quality data can substantially impact strategic decisions. To achieve satisfactory performance using machine learning techniques to predict the rock properties, it is imperative to ensure good data quality before training.

The wells used in this work were drilled with oil-based drilling fluid, which can contaminate the rock samples, considerably affecting the geochemical measurement results, as also reported by Freitas et al. (2022) [35] for presalt rocks. The evaluation of the degree of contamination is a costly task and is traditionally achieved by sample-by-sample evaluation of pyrograms, the results of Rock-Eval pyrolysis, coupled with other geochemical and geological data [37]. Therefore, Algorithm 1, which automatically qualifies the geochemical data and excludes contaminated rock samples, is the first stage of the work flow. Rock-Eval data from 167 rock samples from lacustrine and marine source rocks from different Brazilian sedimentary basins were selected and grouped into two classes (non-contaminated (reliable) or poorly contaminated and contaminated (unreliable)) based on the Rock-Eval product analysis of each rock sample.

As each pyrogram consists of around 1200 data points (one for each time step of the Pyrolysis analysis), we used principal component analysis (PCA) [38,39] to reduce the dimensions of the normalized data to seven components, guaranteeing 99% of the cumulative explained variance. Besides the seven components, the hydrocarbon potential (S_2), the production index (PI), and the number of free hydrocarbons normalized by the TOC (S_1/TOC) were chosen as attributes to train the model after the exploratory analysis (data visualization and analysis). A proportion of 33% of the data was used to evaluate the model's performance. Several machine learning methods were tested, and the random forest method was chosen, as it provided the best classification result on the validation set (Table 1).

Table 1. Metrics for quality control of automated classification of geochemical rock samples into reliable and unreliable.

	PRECISION	RECALL	F1 SCORE
RELIABLE	1.00	0.68	0.81
UNRELIABLE	0.83	1.00	0.91
ACCURACY			0.88
MACRO AVERAGE	0.91	0.84	0.86
WEIGHTED AVERAGE	0.90	0.88	0.87

With the same purpose, Algorithm 2 consists of the evaluation of the quality of the well log data, removing the outliers, mainly based on irregularities in the borehole. It is common for well log data to contain values affected by washouts, as well as other measurements that can be considered outliers for a given work flow. These problematic data need to be removed so that they do not result in misinterpretations in statistical analyses and machine learning flows. The traditional work flow is implemented manually by a geoscientist through the visual evaluation of two-dimensional curves and cross plots. This manual process is very time-consuming, often making it impossible to remove outliers from an extensive database. Accordingly, we developed an automatic outlier removal flow using the DBSCAN (density-based spatial clustering of applications with noise [29]) unsupervised clustering algorithm. The advantage of using DBSCAN for this purpose, compared to other clustering methods, is that in addition to classifying the data into clusters, it also allows for the identification of outliers, that is, points that do not belong to any cluster. The main idea of this application is to use the log data from several different wells so that the machine learning algorithm recognizes the existing patterns in the data and can indicate those measurements that are not part of the expected regular distribution.

3.2. Algorithm 3: 1D Property Estimation

The following step consists of the quantification of the quantity, quality, hydrocarbon potential, and maturity of organic matter using well logs coupled with geochemical data. Traditional methods propose the use of porosity and resistivity logs to estimate the TOC content (e.g., [3,4]). However, it is well known that the physical properties of source rocks allow them to be recognized in other well logs (e.g., [5,6]). Machine learning techniques can help to automatically find relations between those data to quantify not only the TOC but also other geochemical properties (e.g., [40,41]).

The dataset used to train the machine learning models to predict the TOC content and other properties comprises 92 wells. These wells were separated into two distinct datasets: a training dataset, with about 80% of the wells, and a test dataset, with the remainder of the data, acting as a blind test case. The test dataset was used to evaluate the model performance concerning new data, such as the newly drilled well case. The selection of the best model consists of choosing the one with the best performance on the test dataset. The following machine learning algorithms were evaluated: random forest [30], support vector machine (SVR) [42], XGBoost [43], NGBoost [32], and neural networks (multilayer perceptron (MLP)) [44]. TOC and Rock-Eval parameters from the non-contaminated and poorly contaminated cuttings, as well as sidewall and core samples from the Jiquiá source rock from Santos and Campos basins, were selected as targets to perform exploratory data analysis for log selection. At the end of this step, a total of 606 samples were selected, considering only sidewall and core samples due to the high inaccuracy of depths from the cuttings samples. The logs selected as input data to estimate TOC were gamma rays, density, neutron, deep resistivity, compressional sonic, and the burial depth of each well log measurement. These logs were selected due to their known correlation with TOC values [3–6]. The same logs were used for Tmax value estimation, as proposed by Tariq et al. (2020) [40] and Shalaby et al. (2020) [41].

To train these models, the scikit-learn python library was used [45]. Hyperparameter tunning of each algorithm was performed using scikit-learn's grid search with cross valida-

tion tool. This tool allows several parameters to be tested and compared. For each model, the best parameter combination is obtained by selecting the one with the lowest average error on the cross-validation dataset. In this study, 5 cross-validation groups were used. Finally, the best parameter of each algorithm was used to train a model with all the data available in the training dataset. To compare the performance of the models, we evaluated the correlation coefficient between the test data and the results obtained by each model, with Pearson and Spearman correlations explaining linear and non-linear relationships, respectively (Table 2). The F1 score was obtained by considering the TOC quality (poor when TOC < 0.5, fair when $0.5 \leq$ TOC ≤ 1, and good when TOC > 1 [46]) generated by the discretization of the TOC values. The models, in general, showed equivalent performance regarding the correlation between the estimates and the measured laboratory data. However, the NGBoost model showed better performance in terms of F1 score (Table 2). Additionally, besides predicting only a real value, NGBoost incorporates uncertainty estimation through probabilistic prediction, which is the approach whereby the model outputs a full probability distribution [32]. Accordingly, this was the model selected as the best model for TOC and Tmax estimation.

Table 2. Metrics for each applied method.

METHOD	PEARSON	SPEARMAN	F1-SCORE
RANDOM FOREST	0.72	0.77	0.68
SVR	0.66	0.76	0.67
MLP	0.74	0.81	0.68
XGBOOST	0.74	0.81	0.67
NGBOOST	0.72	0.80	0.74

A linear regression between the TOC and the hydrocarbon potential (S_2) measurements was used to obtain the S_2log, and the hydrogen index (HI) was derived from the following relation: $S_2\text{log}/\text{TOClog} \times 100$.

3.3. Algorithm 4: 3D Property Estimation

The low density and velocity characteristic of organic matter allow source rocks to be identified in the seismic data. Therefore, the following step is the 3D estimation of the geochemical properties. Although the application in this work was limited to obtaining a volume of TOC, it can also be applied to estimate other geochemical parameters. As shown in Figure 3, the set of elastic attributes used as an input for this study show a trend similar to that of the TOC well log obtained in the previous step. Hence, the focus of the methodology developed in this work is to capture these relationships according to the fit of a machine learning model relating the properties.

To train the model for 3D prediction of geochemical properties, we used a set of elastic attributes from the seismic inversion as features to obtain the TOC values, as represented by the TOC well log (target) described in Algorithm 3.

As in the 1D property estimation case, NGBoost with decision tree as the base learner was the chosen model. The training data were composed of the continuous TOC well logs (from Algorithm 3) from 3 wells (target), which represent a total amount of 626,372 samples for training, and the respective features were traces of each elastic attribute extracted at the location of the training wells. For validation, we used one well, which corresponds to 17,972 samples (16% of data). After the training step, we applied the trained model to the test data using one blind well with 26,758 samples (25% of data). All wells used in this step are located inside the limits of the 3D seismic data.

Before training, it was necessary to filter the TOC well logs to adapt the frequency of features and the target, since the well logs (target) have a wider-frequency bandwidth than the traces of the seismic inversion attributes (features).

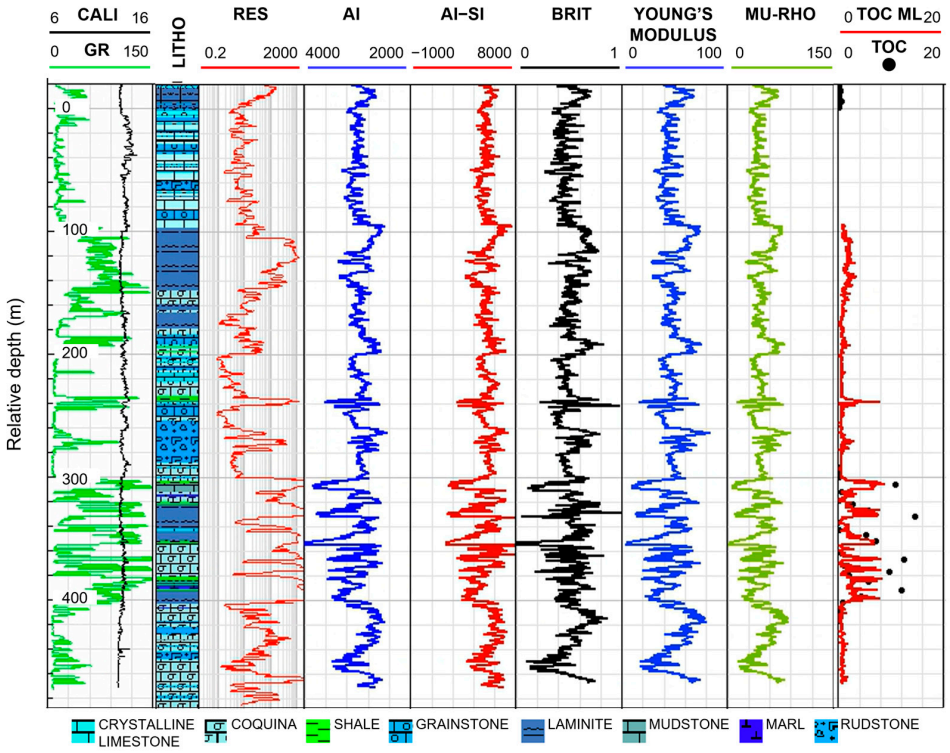

Figure 3. Set of well logs for well 3. From tracks 3 to 7, a set of elastic attributes is exhibited (acoustic impedance, difference between acoustic and shear impedances, brittleness, Young's modulus, and Mu-Rho). On track 8, the TOC log is overlaid with the laboratory measurements of this property. Note the good relation between TOC and all the elastic attributes (higher TOC values correspond to lower elastic attribute values).

4. Results and Discussion

As Algorithm 1 was not yet completed at the time of the development of Algorithm 3, validation of the rock measurements was performed manually. Based on the reliable sidewall and core data, the source rock corresponds predominantly to type I organic matter according to the Langford and Blanc-Valleron (1990) [47] diagram (Figure 4) and has excellent hydrocarbon potential, with an average and maximum TOC of 6% and 36%, respectively, and an average S_2 of 46 up to 358 mg HC/g rock.

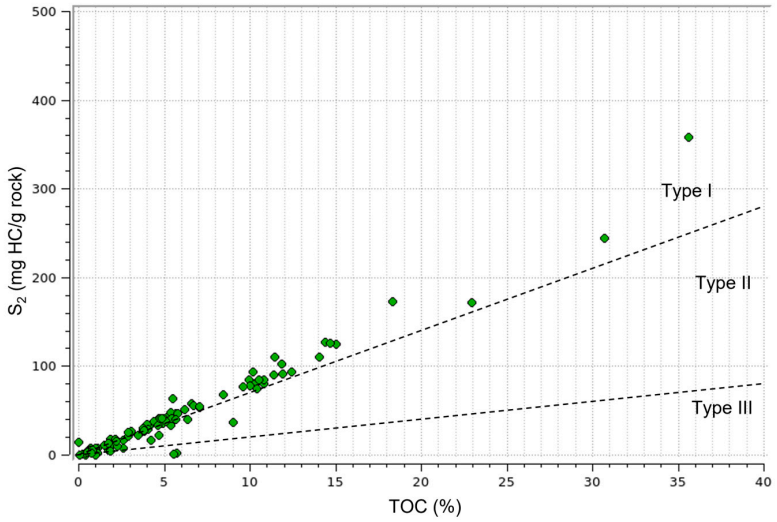

Figure 4. Classification of the source rock measurements using S_2 vs. TOC diagram based on Langford and Blanc-Valleron (1990) [47].

The use of the DBSCAN unsupervised clustering algorithm to recognize and remove outliers from the well-log curves used as features in the TOC prediction flow also presented consistent results. The inputs for the classification were P-sonic, S-sonic, and density well logs. Figures 5 and 6 show the results of the identification of outliers for a given well in the studied area. One can notice that the identified outliers correspond to the higher values of caliper, which suggests that well washout is the cause of the anomalous measurements.

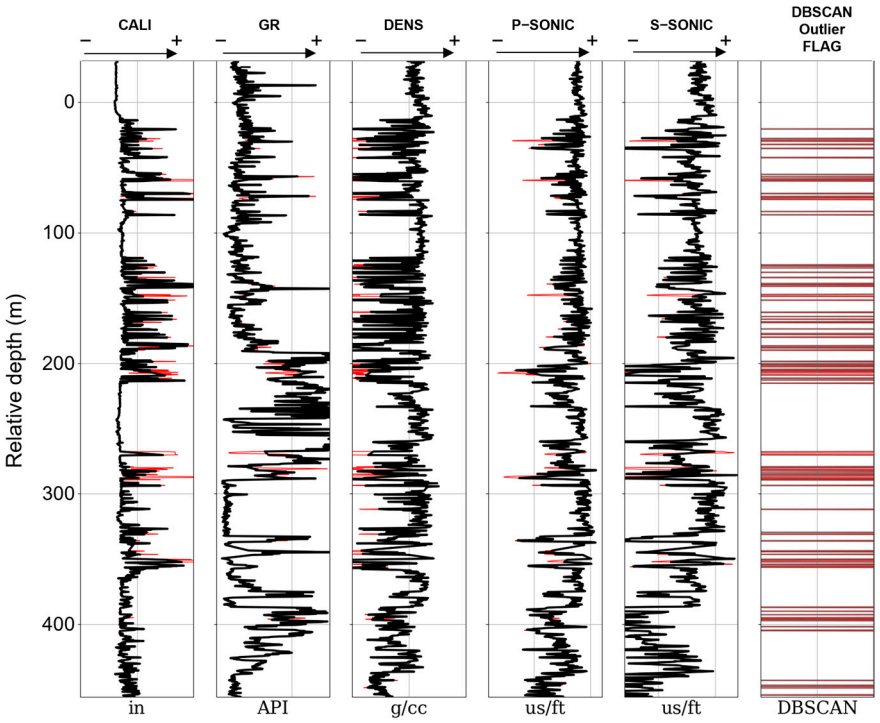

Figure 5. Identification of outliers (red) after performing the DBSCAN algorithm with the following well logs as features: density (DEN), compressional sonic (P-SONIC), and shear sonic (S-SONIC). The outlier flag is shown in the last track. The caliper well log was not used as a feature.

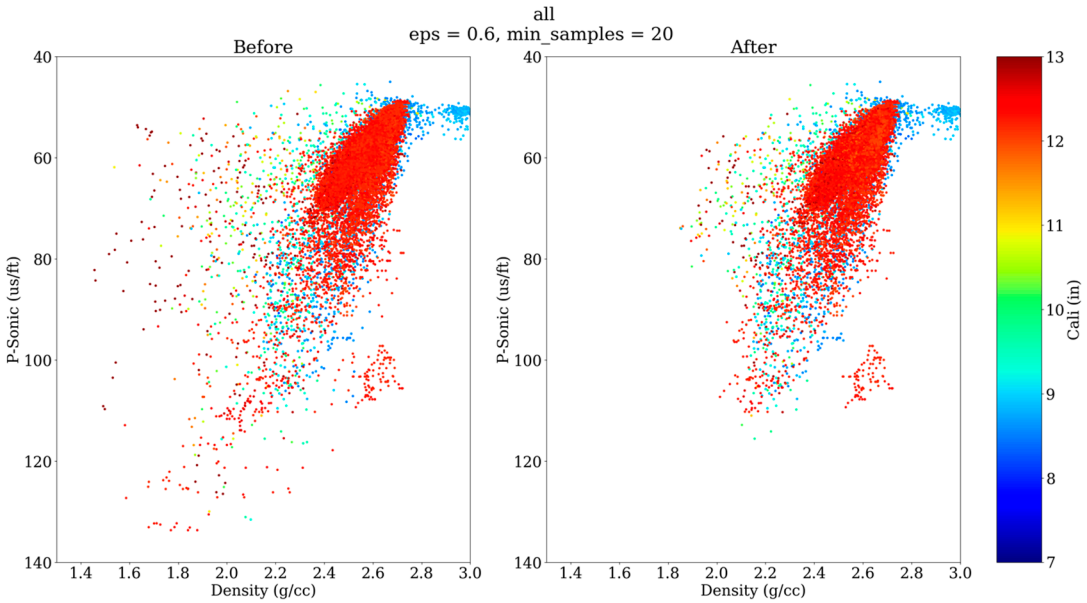

Figure 6. Cross plots of P-sonic as a function of density, colored by caliper before and after the removal of the outliers. Note that the data classified as outlier predominantly correspond to anomalously low densities (lower than 2 g/cc).

The validated rock measurements and well logs were used to train and test the 1D machine learning model, providing good estimations of the geochemical properties continuously in depth, as observed in the blind test result presented in Figure 7 and as indicated in Table 2. One can notice the trend of increasing maturity with increasing depth, as well as the intercalation between carbonates and black shales described by Moreira et al. (2007) [34], where TOC, HI, and S_2 contents tend to be higher in shales, laminites, and mudstones and lower in carbonate reservoir facies.

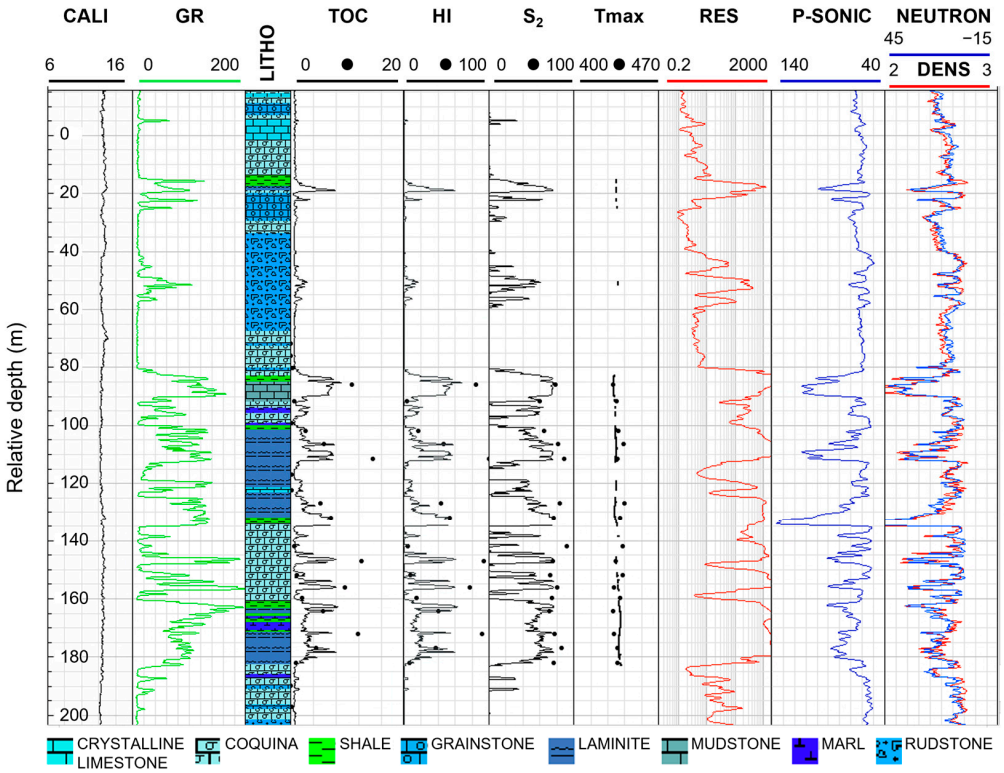

Figure 7. Well logs for well 3: caliper (CAL, in), interpreted lithology (LITHO), Gamma ray (GR, API), resistivity (RES, ohm.m), sonic compressional (P-SONIC, us/ft), density (DENS, g/cm^3), neutron (%); and geochemical logs: TOC (%), HI (mg HC/g TOC), S_2 (mg HC/g rock), and Tmax (°C) for one of the blind test wells, compared with the rock samples measurements (blind test data are represented as black dots, while continuous lines correspond to results obtained by our model).

Different combinations of elastic attributes were tested, aiming for greater accuracy in TOC prediction. Although the importance of the attributes for TOC prediction varies depending on the combination tested in the training, the acoustic impedance (IP) attribute remained with the highest index of importance (Figure 8). The TOC calculated from linear regression with the shear impedance (IS) was used as a benchmark for the prediction (Figure 9), as it is the most conventional approach to estimate TOC volumes from seismic data.

The peak frequency attribute was not used due to its low importance in TOC prediction. Also, we verified that removing the shear impedance (IS) attribute from the list of input attributes did not reduce the prediction accuracy, probably due to redundant attributes in the set generated from simple arithmetic combinations of IS with others. Figure 10 shows a blind test for the model whose metrics are indicated to be the best choice of parameters and attributes as features for TOC prediction. Note that the machine-learning-predicted TOC log (red) is very similar to that used as the target for the prediction (blue), whereas the regression with the IS attribute alone (green) does not provide a satisfactory prediction of this property. The values of Pearson's correlations between the predicted well logs and the target, also shown in Figure 10, validate the greater accuracy of NGBoost prediction. In addition to greater accuracy in prediction using machine learning compared to the benchmark, another advantage of using NGBoost is predicting parameters of a

normal statistical distribution, allowing for the estimation and analysis of uncertainty in the predictions. The red line represented as the TOC prediction using NGBoost in Figure 10 is the P50 of the prediction, and the highlighted red space around the curve is the interval between P10 (pessimistic) and P90 (optimistic) predictions.

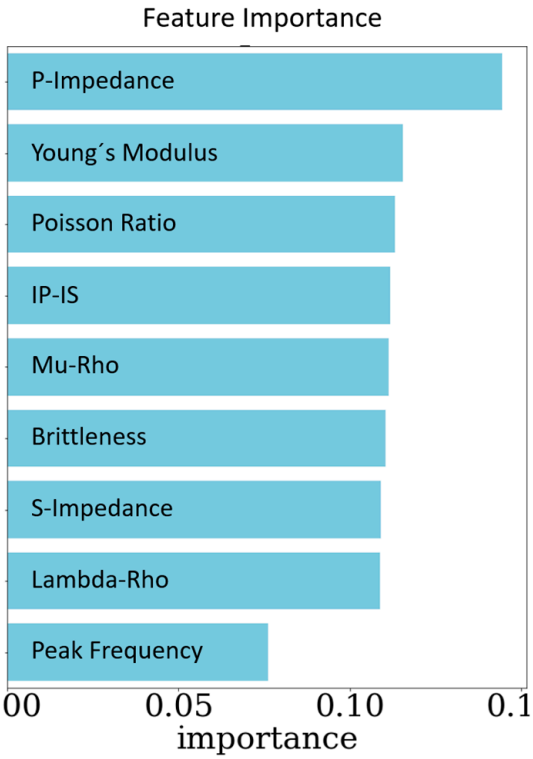

Figure 8. Feature importance of tested attributes. P-impedance always shows the highest value of importance. Among all nine tested attributes, peak frequency and S-impedance were discarded for the final training set.

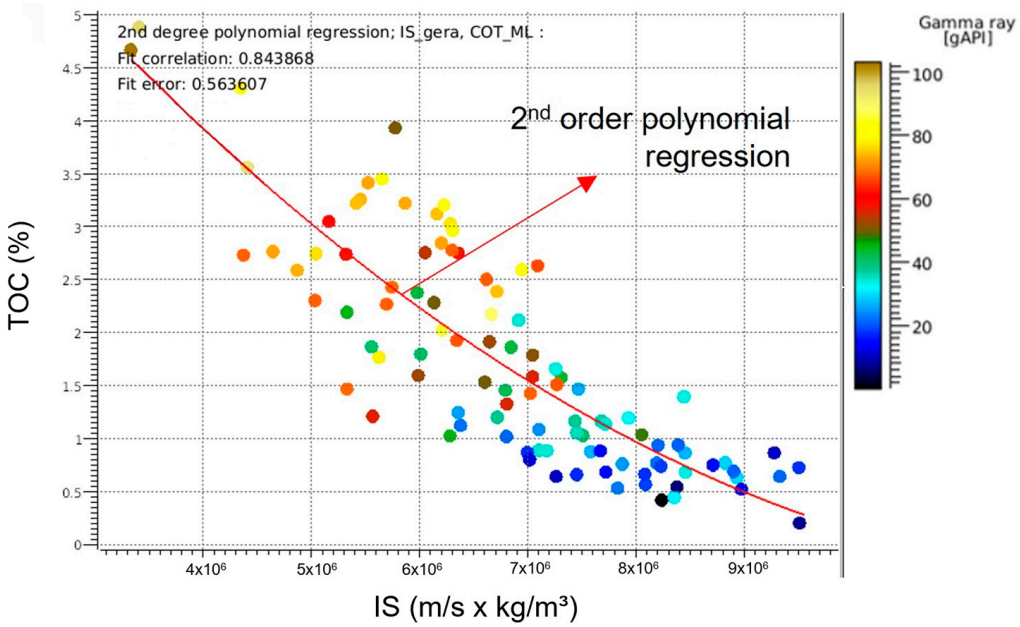

Figure 9. TOC as a function of S-impedance (IS). The linear regression is calculated as $TOC = 8.57 - 1.37 \times 10^{-6} \times IS + 5.31 \times 10^{-14} \times IS^2$.

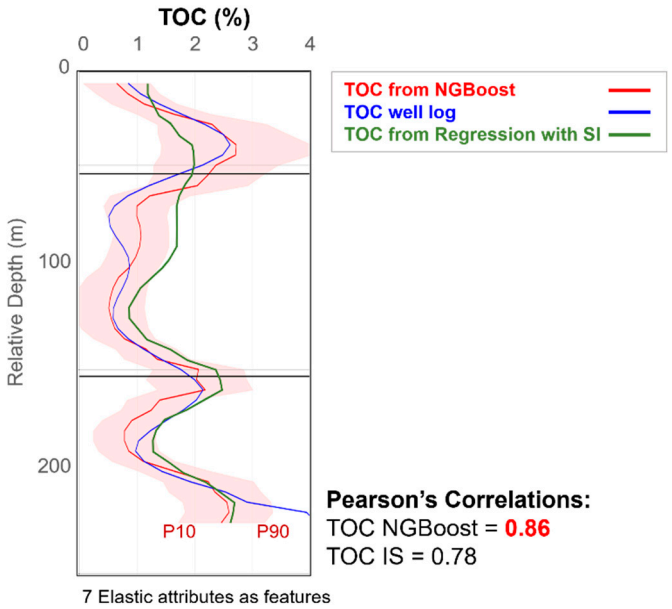

Figure 10. The plot above shows the results of the total organic carbon (TOC) prediction. The red strip is the range of possible TOC values between the P10 and P90 quantiles. The blue line is the well-log TOC (target), the red line is the NGBoost prediction (or P50), and the gray line is the TOC estimate from the linear regression with the single S-impedance attribute. The TOC calculated only with the S-impedance attribute (green line) shows a lower correlation with the target than that generated using machine learning (highlighted in red on the right side of the figure) in all tested scenarios.

Finally, based on the performed tests, the attributes chosen to generate the final trained model were P-impedance, brittleness, Young's modulus, Lambda-Rho, Mu-Rho, Poisson ratio, and the difference between P- and S-impedance. This trained model was used for the 3D prediction of TOC, as described below.

The vertical resolution of the input seismic data for this study is 105 m, and the inverted data (used as a feature for TOC prediction) have a resolution of 80 m. The inversion resolution gain is due to the deconvolution of the seismic pulse intrinsic to the inversion process. When the inversion results are used as input to machine learning models, which are complex and non-linear, this increase in vertical resolution is further enhanced, which translates into a better definition of the top and bottom of the richest layers in OM. Figure 11 shows a cross section from the P50 predicted TOC volume compared with the target well logs. One can observe a large vertical and lateral variability of TOC content on the source rock, alternating between rich and poor organic matter intervals, possibly due to the intercalation of carbonates and black shales described by Moreira et al. (2007) [34] and observed in the wells. The TOC well logs are shown in Figure 11 (black wiggle) superimposed on an arbitrary cross section of the TOC volume that crosses the wells. Note that the behavior of the TOC log agrees with that of the volumetric estimate of this property, evidencing the good fit between the logs (also generated via ML) and the predicted volume. The property contrasts observed in the wells can be traced laterally. Owing to the good lateral continuity of the estimated volume using ML techniques, the interpretation of the top and base of the main source rock intervals becomes a much easier task (Figure 12). After analyzing the TOC volumes generated in this study, the intervals with the highest TOC became more evident. Figure 13 depicts another cross section of the TOC volume where there is an indication of intervals rich in organic matter. Two of those intervals appear to be present only inside grabens (e.g., in the structural low in Figure 13, where the intervals highlighted by the yellow arrows seem to pinch out to the right-hand side of the figure, against the structural high) and possibly were not yet identified in any drilled wells. That is probably due to those intervals not being deposited over the adjacent structural highs (where there was a shallower paleobathymetry) or being eroded there. As well

locations were selected with a focus on exploring better-quality carbonate reservoirs (thus, in shallower paleobathymetries), only the organically rich intervals also deposited on top of structural highs can be identified in wells. In our example, some of these organically rich intervals interpreted from our results were sampled by wells drilled in the area, but they do not appear as such in conventional seismic data. That is possibly due to the lower thickness of those potential source rock intervals relative to the original seismic data resolution, while our TOC volume has a higher resolution than that of the original seismic data, as mentioned above. Thus, TOC volumes can also be used to complement seismic and well data as indicators of potential source rock intervals.

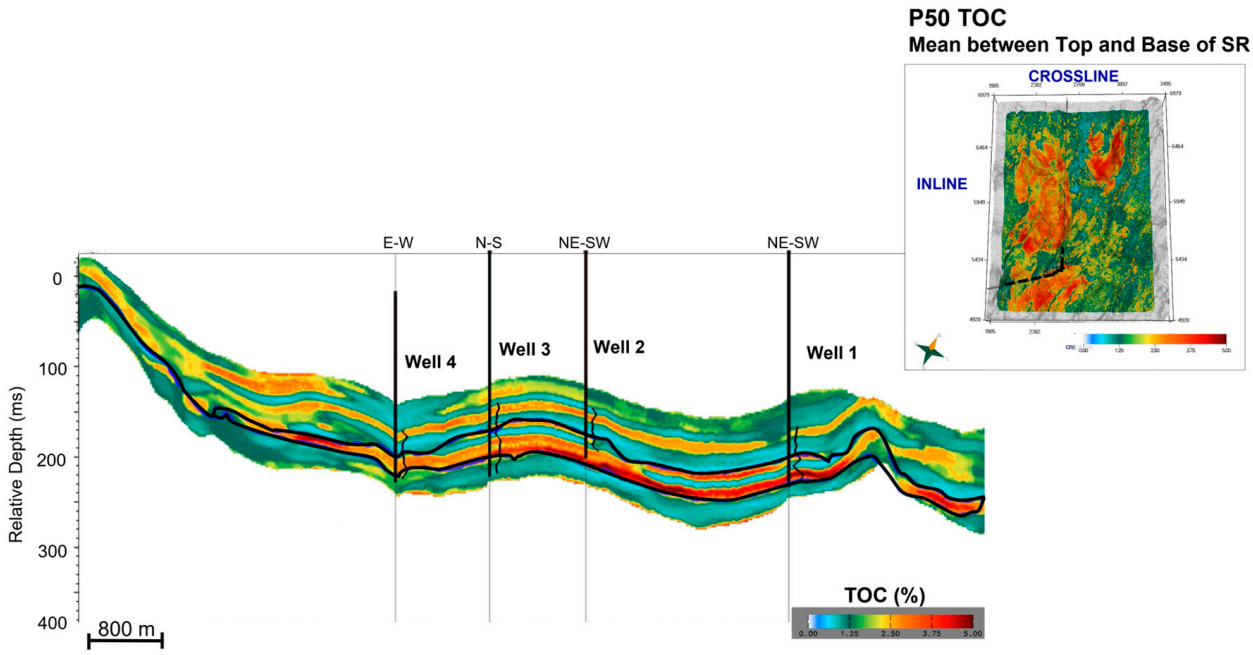

Figure 11. TOC well logs displayed on the same scale as seismic volumes. Note the good fit with the well logs and the vertical resolution gain provided by the ML application. The blue lines are the interpreted top and base of the main source rock (SR) interval.

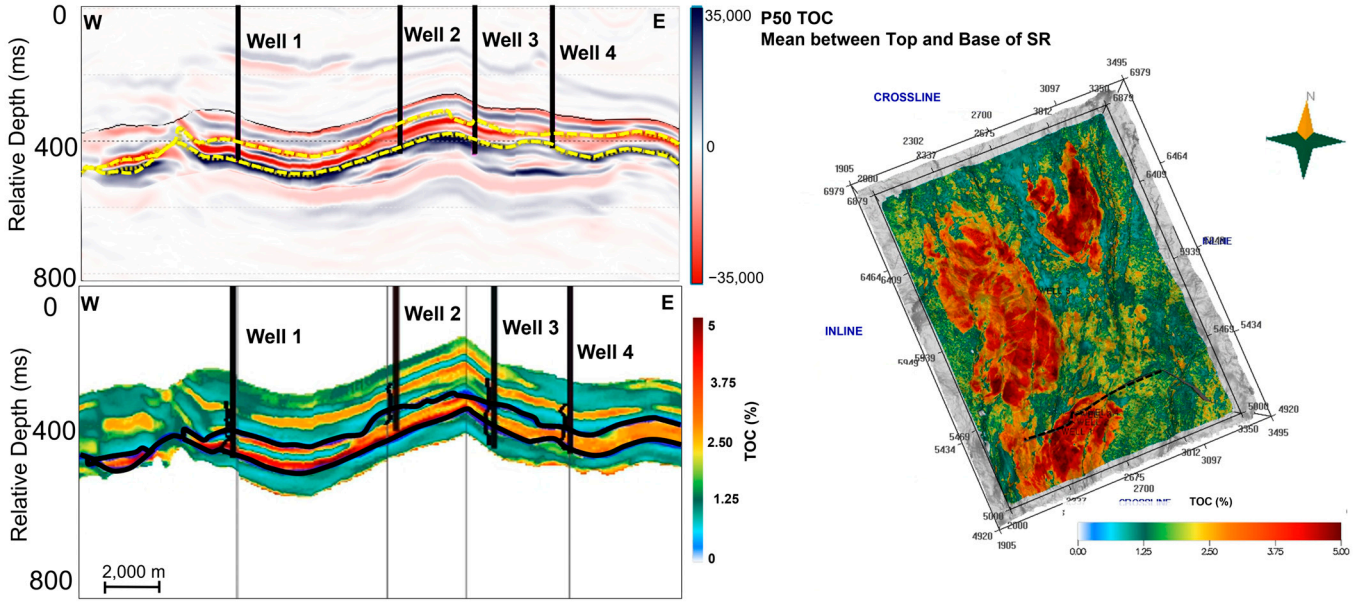

Figure 12. Seismic sections showing the seismic amplitude (above) and predicted TOC (below) with the interpreted top and base of the source rock (yellow dashed lines above and blue lines below).

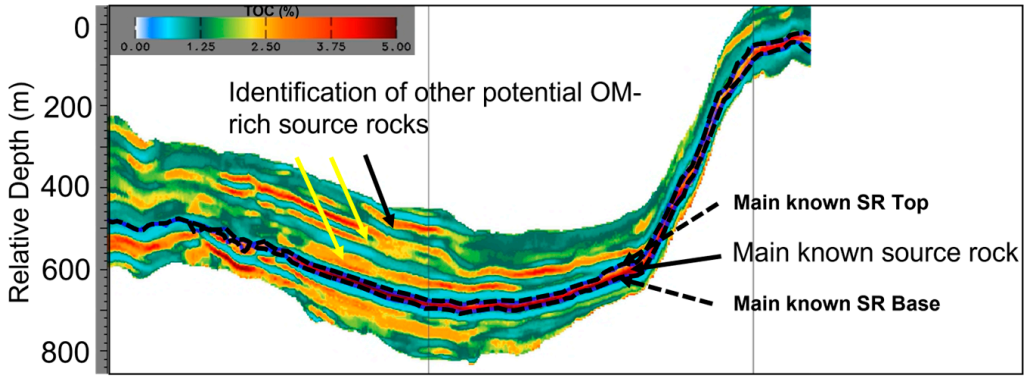

Figure 13. Seismic section showing the predicted TOC volume with the interpreted top and base of the main source rock (SR; black dashed lines). In addition to allowing for tracking of known organically rich source rock intervals, the TOC volume can also indicate new intervals presenting potential source rocks.

Figure 14 displays a map of the average TOC (P50) in the interval between the top and base of the source rock, showing the spatial variations of this property. It can be observed that TOC tends to increase from high structure blocks towards the lows in the northern and western portions of the studied area. However, in the eastern portion, the richness of the source rock is relatively lower. At shallower depths, the source rock was not deposited or was eroded. As the model provides a volumetric probabilistic result, it is possible to obtain a statistical distribution of the predicted property and to extract the information through maps and sections. Figure 15, for example, shows geobodies with high TOC values (higher than 2.5%) for optimistic (P90), pessimistic (P10), and realistic (P50) scenarios. One can observe that even in the pessimistic scenario, there are high values of TOC in the grabens on the north and west, and the eastern area represents the higher charge risk. All the information revealed from the application of this methodology is available for petroleum system analysis assessment and to be incorporated in the numerical basin modeling to reduce uncertainties in the quantification of exploration risks.

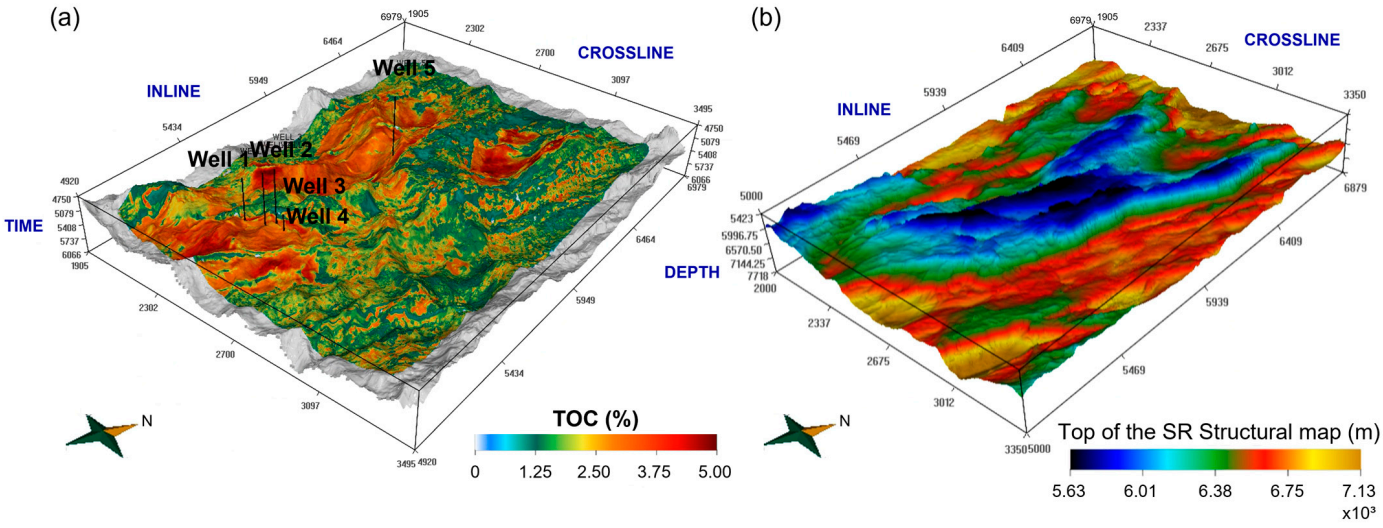

Figure 14. Average TOC map extracted between the top and the base of the source rock interval (**a**) and depth map of the base of the source rock (**b**).

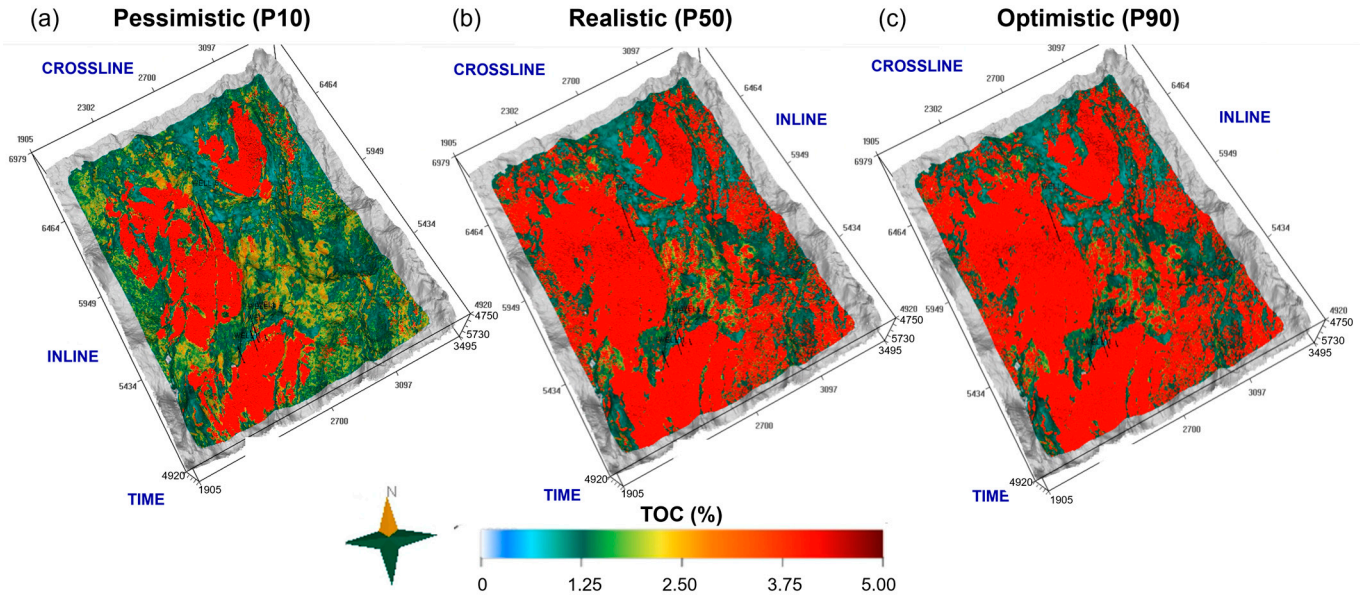

Figure 15. Geobodies of high TOC values (higher than 2.5%, in red) for pessimistic (**a**), realistic (**b**), and optimistic (**c**) scenarios.

5. Conclusions

We propose an integrated work flow based on machine learning methods to characterize source rocks from rock to seismic. Despite being a pilot project and the need to test it in other areas, as shown in the discussion of the results, the methodology provided satisfactory results. It improved the evaluation of input data quality and the estimation of geochemical properties when laboratory data are scarce or absent, ensuring the use of only reliable information and the integration of different datasets at different scales. The use of multiple attributes to estimate TOC combined with the application of machine learning techniques allowed for the estimation of TOC volumes with higher precision and resolution than the input seismic volumes. Once the machine learning models are trained, it is possible to predict 1D and 3D geochemical properties in real time. Therefore, it provided more robust results in a reduced time compared to traditional approaches.

The application of the methodology in the Jiquiá interval suggests that in the studied area, higher TOC values are located in the northern and western grabens, and the eastern area presents a higher charge risk. At the structural high, the source rock was not deposited or eroded. The reported results can be used as input in petroleum systems analysis and are important to mitigate charge risks and fluid assessments in exploration prospects. The interpretation of these results must be deeply allied to the geological knowledge of the area, since, even using sophisticated techniques of multiattribute predictions, ambiguities remain in shales poor in organic matter and reservoirs. We intend to continue this work by applying the methodology to study the Jiquiá source rock in similar areas, as well as to study other source rocks, possibly applying transfer learning techniques.

Author Contributions: Conceptualization, M.A.A.d.A.d.R. and A.C.D.; Methodology, M.A.A.d.A.d.R., A.C.D., C.E.D.R., A.L.K., A.M.d.S., E.d.S.P. and V.G.S.; Validation, M.A.A.d.A.d.R.; Data Curation, M.A.A.d.A.d.R., A.C.D., A.M.d.S. and V.G.S.; Formal Analysis, M.A.A.d.A.d.R., A.C.D., C.E.D.R., A.L.K., A.M.d.S., E.d.S.P. and V.G.S.; Software, C.E.D.R., A.L.K. and A.M.d.S.; Visualization, M.A.A.d.A.d.R., A.C.D., C.E.D.R., A.L.K., A.M.d.S., E.d.S.P. and V.G.S.; Writing—Original Draft Preparation, M.A.A.d.A.d.R., A.C.D., C.E.D.R., A.L.K., A.M.d.S., E.d.S.P. and V.G.S.; Writing—Review and Editing, M.A.A.d.A.d.R.; Supervision, M.A.A.d.A.d.R. and A.C.D.; Project Administration, M.A.A.d.A.d.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Petrobras.

Data Availability Statement: Data associated with this research are confidential and cannot be released.

Acknowledgments: The authors thank Petrobras for making this project possible and for allowing the results to be published. The authors also thank their colleagues, Luílson Leal and Sá, Sofia de Abreu e Lima Correia, Taíssa Rego Menezes, Regina Binotto, and Erick Costa e Silva Talarico for their input and support provided for the execution of this work.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Tissot, B.P.; Welte, D.H. *Petroleum Formation and Occurrence. A New Approach to Oil and Gas Exploration*; Springer: New York, NY, USA, 1978; 538p.
2. Espitalié, J.; Laporte, J.L.; Madec, M.; Marquis, F.; Leplat, P.; Paulet, J.; Boutefeu, A. Méthode rapide de caractérisation des roches mères, de leur potentiel pétrolier et de leur degré d'évolution. *Rev. De L'institut Français Du Pétrole* **1977**, *32*, 23–42. [CrossRef]
3. Passey, Q.R.; Creaney, S.; Kulla, J.B.; Moretti, F.J.; Stroud, J.D. A practical model for organic richness from porosity and resistivity logs. *AAPG Bull.* **1990**, *74*, 1777–1794.
4. Carpentier, B.; Huc, A.Y.; Bessereau, G. Wireline logging and source rocks estimation of organic carbon by the CARBOLOG method. *Log Anal.* **1991**, *32*, 279–297.
5. Fertl, W.H.; Chilingarian, G.V.; Yen, T. Organic Carbon Content and Source Rock Identification Based on Geophysical Well Logs. *Energy Sources* **1986**, *8*, 381–439. [CrossRef]
6. Sun, S.Z.; Sun, Y.; Sun, C.; Liu, Z.; Dong, N. Method of calculating total organic carbon from well logs and its application on rock's properties analysis. In Proceedings of the GeoConvention 2013: Integration, Calgary, AB, Canada, 6–10 May 2013.
7. Carcione, J.M. AVO effects of a hydrocarbon source-rock layer. *Geophysics* **2001**, *66*, 419–427. [CrossRef]
8. Løseth, H.; Wensaas, L.; Gading, M.; Duffaut, K.; Springer, M. Can hydrocarbon source rocks be identified on seismic data? *Geol. Soc. Am.* **2001**, *39*, 1167–1170. [CrossRef]
9. Jia, J.; Liu, Z.; Meng, Q.; Liu, R.; Sun, P.; Chen, Y. Quantitative evaluation of oil shale based on well log and 3-D seismic technique in the Songliao Basin, northeast China. *Oil Shale* **2012**, *29*, 128–150. [CrossRef]
10. Gading, M.; Wensaas, L.; Løseth, H. Source rocks from seismic, Part-2—Applications. In Proceedings of the Eage Conference & Exhibition Incorporating, Copenhagen, Denmark, 4–7 June 2012.
11. Yenugu, M.; Han, D. Seismic characterization of kerogen maturity: An example from Bakken shale. In Proceedings of the Society of Exploration Geophysicists International Exposition and Annual Meeting, Houston, TX, USA, 22–27 September 2013.
12. Sharma, R.K.; Chopra, S.; Vernengo, L.; Trinchero, E.; Sylwan, C. Reducing uncertainty in characterization of Vaca Muerta Formation Shale with poststack seismic data. *Lead. Edge* **2015**, *34*, 1462–1467. [CrossRef]
13. Ouadfeul, S.A.; Aliouane, L. Total organic carbon estimation in shale-gas reservoirs using seismic genetic inversion with an example from the Barnett Shale. *Lead. Edge* **2016**, *35*, 790–794. [CrossRef]
14. Mraz, T.; Dubow, J.; Rajeshwar, K. Acoustic wave propagation in oil shale: 1. Experiments. *Fuel* **1983**, *62*, 1215–1222. [CrossRef]
15. Vernik, L.; Nur, A. Ultrasonic velocity and anisotropy of hydrocarbon source rocks. *Geophysics* **1992**, *57*, 727–735.
16. Vernik, L.; Liu, X. Velocity anisotropy in shales: A petrophysical study. *Geophysics* **1997**, *62*, 521–532. [CrossRef]
17. Carcione, J.M. A model for seismic velocity and attenuation in petroleum source rocks. *Geophysics* **2000**, *65*, 1080–1092.
18. Eseme, E.; Urai, J.L.; Krooss, B.M.; Littke, R. Review of mechanical properties of oil shales: Implications for exploitation and basin modelling. *Oil Shale* **2007**, *24*, 159. [CrossRef]
19. Zhu, Y.; Liu, E.; Martinez, A.; Payne, M.A.; Harris, C.E. Understanding geophysical responses of shale-gas plays. *Lead. Edge* **2011**, *30*, 332–338. [CrossRef]
20. Vernik, L.; Milovac, J. Rock physics of organic shales. *Lead. Edge* **2011**, *30*, 318–323. [CrossRef]
21. Qin, X. Vp-Vs Relations of Organic-Rich Shales. Master's Thesis, Faculty of the Department of Earth and Atmospheric Sciences, University of Houston, Houston, TX, USA, 2013.
22. Qian, Z. Geophysical Responses of Organic-Rich Shale and the Effect of Mineralogy. Master's Thesis, Faculty of the Department of Earth and Atmospheric Sciences, University of Houston, Houston, TX, USA, 2013.
23. Li, W.; Zhang, Z.; Li, Y. Some aspects of excellent marine source rock formation: Implications on enrichment regularity of organic matter in continental margin basins. *Chin. J. Geochem.* **2015**, *34*, 47–54. [CrossRef]
24. Sayers, C.M.; Fisher, K.; Walsh, J.J. Sensitivity of P-and S-impedance to the presence of kerogen in the Eagle Ford Shale. *Lead. Edge* **2015**, *34*, 1482–1486. [CrossRef]
25. Sayers, C.M. The effect of kerogen on the AVO response of organic-rich shales. *Lead. Edge* **2013**, *32*, 1514–1519. [CrossRef]
26. Sengupta, M.; Jacobi, D.; Eichmann, S.; Wallet, B.; Altowairqi, Y.; Alsinan, S. Seismic assessment of maturity and richness in carbonate source rocks. In Proceedings of the International Meeting for Applied Geoscience & Energy, Houston, TX, USA, 28 August–2 September 2022.
27. Del Monte, A.A.; Antonielli, E.; De Tomasi, V.; Luchetti, G.; Paparozzi, E.; Gambacorta, G. Methods for source rock identification on seismic data: An example from the Tanezzuft Formation (Tunisia). *Mar. Pet. Geol.* **2018**, *91*, 108–124. [CrossRef]
28. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016. Available online: https://www.deeplearningbook.org (accessed on 2 May 2021).

29. Ij, H. Statistics versus machine learning. *Nat. Methods* **2018**, *15*, 233.
30. Ho, T.K. Random Decision Forests. In Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, Canada, 14–16 August 1995.
31. Ester, M.; Kriegel, H.; Sander, J.; Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, Portland, OR, USA, 2–4 August 1996.
32. Duan, T.; Avati, A.; Ding, D.Y.; Thai, K.K.; Basu, S.; Ng, A.Y.; Schuler, A. NGBoost: Natural Gradient Boosting for Probabilistic Prediction. *arXiv* **2019**, arXiv:1910.03225. Available online: https://arxiv.org/abs/1910.03225 (accessed on 2 May 2021).
33. Estrella, G.; Mello, M.R.; Gaglianone, P.C.; Azevedo, R.L.M.; Tsubone, K.; Rossetti, E.; Concha, J.; Bruning, I.M.R.A. The Espirito Santo Basin (Brazil) source rock characterization and petroleum habitat. *Am. Assoc. Pet. Geol. Bull.* **1984**, *35*, 253–271.
34. Moreira, J.L.P.; Madeira, C.V.; Gil, J.A.; Machado, M.A.P. Bacia de Santos. *Bol. De Geociências Da Petrobras* **2007**, *15*, 531–549.
35. Freitas, V.A.; Vital, J.C.S.; Rodrigues, B.R.; Rodrigues, R. Source rock potential, main depocenters, and CO_2 occurrence in the pre-salt section of Santos Basin, southeast Brazil. *J. S. Am. Earth Sci.* **2022**, *115*, 103760. [CrossRef]
36. Damasceno, A.C.; Korenchendler, A.L.; Da Silva, A.M.; Da Silva Praxedes, E.; De Almeida Dos Reis, M.A.A.; Silva, V.G. Source rock evaluation from rock to seismic: Integrated machine learning based workflow. In Proceedings of the IMAGE, Houston, TX, USA, 28 August–1 September 2022.
37. Peters, K.E. Guidelines for Evaluating Petroleum Source Rock Using Programmed Pyrolysis. *Am. Assoc. Pet. Geol. Bull.* **1986**, *70*, 318–329.
38. Pearson, K. On lines and planes of closest fit to systems of points in space. *Lond. Edinb. Dublin Philos. Mag. J. Sci.* **1901**, *2*, 559–572. [CrossRef]
39. Hotelling, H. Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* **1933**, *24*, 417–441. [CrossRef]
40. Tariq, Z.; Mahmoud, M.; Abouelresh, M.; Abdulraheem, A. Data-Driven Approaches to Predict Thermal Maturity Indices of Organic Matter Using Artificial Neural Networks. *Am. Chem. Soc.* **2020**, *5*, 26169–26181. [CrossRef]
41. Shalaby, M.R.; Malik, O.A.; Lai, D.; Jumat, N.; Islam, M. Thermal maturity and TOC prediction using machine learning techniques: Case study from the Cretaceous–Paleocene source rock, Taranaki Basin, New Zealand. *J. Pet. Explor. Prod. Technol.* **2020**, *10*, 2175–2193. [CrossRef]
42. Cortes, C.; Vapnik, V. Support vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
43. Chen, T.; Guestring, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
44. Rumelhart, D.E.; Hinton, G.E.; Williams, R.J. Learning representations by back-propagating errors. *Nature* **1986**, *323*, 533–536. [CrossRef]
45. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
46. Peters, K.E.; Cassa, M.R. Applied source rock geochemistry. In *The Petroleum System—From Source to Trap*; Magoon, L.B., Dow, W.G., Eds.; American Association of Petroleum Geologists: Tulsa, OK, USA, 1994; Volume 60, pp. 93–120. [CrossRef]
47. Langford, F.F.; Blanc-Valleron, M.M. Interpreting Rock-Eval pyrolysis data using graphs of pyrolyzable hydrocarbons versus total organic carbon. *Am. Assoc. Pet. Geol. Bull.* **1990**, *74*, 799–804. [CrossRef]