









A Nondestructive Alternative for Kerogen Type Determination in Potential Hydrocarbon Source Rocks Using Hyperspectral Data and Machine Learning

Tainá T. Guimarães , Lucas S. Kupssinskü , Milena B. Cardoso , Leonardo Bachi , Alysson S. Aires , Emilie C. Koste, André L. D. Spigolon , Luiz Gonzaga Jr. , *Member, IEEE*, and Mauricio R. Veronez 

Abstract—In petroleum geology, to assess the hydrocarbon generation potential in source rocks involves the determination of the kerogen type by some destructive method. The usage of such methods is a bottleneck in the process because it is time-consuming, requires specialized tools and personnel, and ends up destroying the rock sample, so it is not possible to do any posterior analysis. This study presents an alternative method for determination of the kerogen type that is fast and nondestructive using hyperspectral data and machine learning techniques. The method is validated using five distinct supervised learning algorithms that were applied to spectral data collected in rock samples from Taubaté Basin, Brazil, of an outcrop whose rocks have a wide range of hydrocarbon generation potential. Cores and samples were collected from the outcrop and had their kerogen type determined by geochemical analyses performed in the laboratory. The robustness of the method is evaluated in two distinct experiments. In the first one, the hyperspectral dataset was collected using a nonimaging spectroradiometer; in the second one, the method uses nonimaging hyperspectral data as training and is tested in hyperspectral images collected. In both experiments, the method was able to establish a relationship between selected spectral features and the kerogen type of the source rocks sampled. The results obtained in this article are prospective for nondestructive classification of kerogen type (and, consequently, the hydrocarbon generation potential) since most of the models generated achieved accuracy above 0.8 in the validation step and 0.75 in the test step.

Index Terms—Classification, hydrocarbon source rock, hyperspectral, kerogen type, machine learning (ML).

Manuscript received 29 April 2022; revised 15 July 2022; accepted 19 July 2022. Date of publication 29 July 2022; date of current version 17 August 2022. This work was supported in part by Petróleo Brasileiro S.A. (PETROBRAS) and in part by Agência Nacional do Petróleo, Gás Natural e Biocombustíveis under Grant 4600556376 and Grant 4600583791. (Corresponding authors: Tainá T. Guimarães; Mauricio R. Veronez.)

Tainá T. Guimarães, Lucas S. Kupssinskü, Milena B. Cardoso, Leonardo Bachi, Alysson S. Aires, Emilie C. Koste, Luiz Gonzaga Jr., and Mauricio R. Veronez are with the Graduate Program in Applied Computing and the Vizlab—X-Reality and Geoinformatics Lab, Vale do Rio dos Sinos University, São Leopoldo CEP 93022-750, Brazil (e-mail: tainat@edu.unisinos.br; lkupssinsku@edu.unisinos.br; milenabcar@edu.unisinos.br; emiliek@edu.unisinos.br; alyssonas@unisinos.br; emiliek@edu.unisinos.br; lgonzaga@unisinos.br; veronez@unisinos.br).

André L. D. Spigolon is with the Research and Development Center of Petrobras, Rio de Janeiro CEP 29941-915, Brazil (e-mail: andrespigolon@petrobras.com.br).

This article has supplementary downloadable material available at <https://doi.org/10.1109/JSTARS.2022.3195088>, provided by the authors.

Digital Object Identifier 10.1109/JSTARS.2022.3195088

I. INTRODUCTION

IN PETROLEUM geology, rocks that are able to generate hydrocarbons, whether in the form of oil or gas, are called source rocks [1]. These rocks are usually composed of sedimentary material of very fine granulometry and contain organic matter in quantity and quality sufficient that, under ideal conditions, the kerogen content is degraded for hydrocarbon generation [2], [3]. The characterization of this rock type is often performed by analyses based on the organic matter of the rock, both in relation to its quantity (organic richness) and quality (kerogen type), its generation potential, and thermal maturity [1], [4], [5]. In this sense, Rock-Eval pyrolysis is a valuable geochemical method as it allows us to obtain the quality of organic matter, thermal maturity, and generation potential [3]. The determination of the kerogen type in the source rock is performed using data obtained by Rock-Eval pyrolysis and also by the amount of organic matter, measured as total organic carbon (TOC), which are later evaluated in the Van Krevelen-type diagram [6].

Although the geochemical data mentioned provide valuable information for geoscientists, these are obtained from laboratory analysis, which are often destructive or requires sample processing, besides involving high costs and skilled labor [7], [8]. Furthermore, these data suffer from a spatial-scale limitation [8], [9] as they are often obtained from point sampling, whether samples collected in outcrops or cores.

A nondestructive and fast alternative to overcome the limitations of traditional geochemical methods is by using hyperspectral remote sensing data to estimate the geochemical parameters through methods that are based on reflectance spectroscopy [8], [10], [11]. Reflectance is the fraction of the light intensity that is reflected by a target [12] and is usually presented by a curve in percentage values to different wavelengths, called a spectral signature. Reflectance spectroscopy is widely exploited to collect compositional information from rocks in a nondestructive and replicable way [13]. This is possible because processes that occur on a molecular scale in rock cause light absorbance in some specific wavelengths [14]. As an example, organic carbon compounds (hydrocarbons) causes spectral features around 1700 and 2300 nm [7], [8], [13], [15].

Hyperspectral data collected by nonimaging sensors with a high spectral resolution, such as spectroradiometers, are extremely valuable for creating spectral libraries, analysis of spectral signatures, and their relationship with the compounds in rocks. However, given that the purpose of using hyperspectral remote sensing is not only to detect these characteristics or to infer them but also to map their spatial distribution, reflectance spectra for each spatial point of a sample or outcrop must be acquired [8]. This is possible using an imaging sensor that, combining reflectance spectroscopy with a high-resolution digital image, provides a reflectance spectrum collected for each pixel of the image [8], [13], [14], [16]. Therefore, by allowing a synoptic view of large sections in spatial detail, hyperspectral images can be used to guide sample collection on outcrops [8], [13], [17]. In addition, they can be used on the mapping of samples and cores as a support for the selection of relevant samples for more detailed analyses, which should reduce the number of samples needed for the investigation, and also on helping with the lithological description process [9], [13], [18].

Hyperspectral data analysis in source rocks is a complex task due to the mixture of minerals and, consequently, its spectral signatures. Robust and automated techniques are needed to identify patterns in slight variations of the spectra. Recently, the use of machine learning (ML) techniques has been explored for the analysis of hyperspectral data as it can perform fast and objective analyses, improving the accuracy and robustness of the models created compared to traditional methods [18]–[21].

Therefore, in this article, we explored ML techniques to classify potential source rocks according to their kerogen type from hyperspectral data. Samples of a sedimentary basin with high hydrocarbon generation potential (Tremembé Formation, Taubaté Basin, Southeastern Brazil) were used to assess the performance of the selected ML algorithms: logistic regression (LR), k-nearest neighbors (KNN), random forest (RF), support vector machine (SVM), and multilayer perceptron (MLP). The algorithms were trained and validated with data collected from a spectroradiometer (nonimaging hyperspectral sensor) as input and the kerogen type as output. Two experiments were performed: the first considering two drill-cores collected in the same study area, one to train and validate the models and another to test them; and the second explored data from hand samples collected at the outcrop and hyperspectral images of samples as a test dataset.

II. METHODOLOGY

The methodology we adopted for this study is presented in the flowchart in Fig. 1. The main steps in the execution of the proposed method are represented in the flowchart by letters to facilitate their reference throughout the manuscript.

Overall, samples of potential source rocks from the study area (A) had their geochemical and hyperspectral data collected (B) and preprocessed (C) for ML experiments to be carried out (D). Although we performed two experiments, both included the same classifier algorithms, features, and approaches of training (E), hyperparameter selection (F), and validation (G). The difference between them is in the data used in the model testing

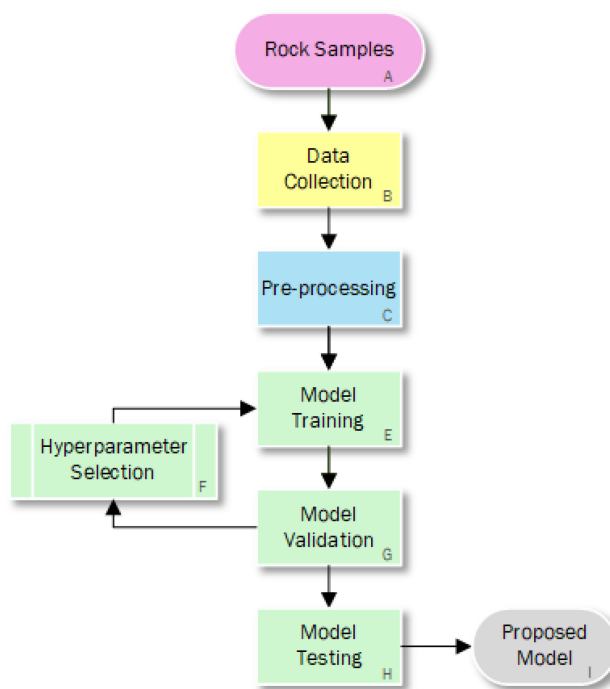


Fig. 1. Simplified flowchart of this study. Each color represents a macrostep of the process.

(H) and, consequently, in their results and discussions. Each of these steps has been detailed in the following sections.

A. Study Area and Sampling

The study was performed in a mining area located in the city of Tremembé, state of São Paulo, southeastern Brasil, which is observed a vertical outcrop [see Fig. 2(a)]. In this place, rock samples were collected from a representative section of the Tremembé Formation (Taubaté Basin) containing rocks with high potential hydrocarbons source.

The mapped outcrop represents the Tremembé Formation, a playa-lake-type lacustrine system, from oligocenic age, located in the central portion of the Taubaté Basin, a part of the Continental Rift of Southeastern Brazil [22]. The Taubaté basin was filled with alluvial deposits intertwined with lacustrine deposits as a consequence of the alteration of tectonic and sedimentation rates influenced by climatic oscillations [22]. The Tremembé Formation is predominantly composed of a succession of sedimentary rocks composed of massive green claystones, rhythms of bituminous shales and marls, dolomites, and sandstones associated with lake and swamp deposits [22]. Rocks from this formation may have high TOC contents, close to 30% in some portions [23], [24], indicating a high hydrocarbon generation potential, being evaluated as analogous to other Brazilian source rocks in studies of hydrocarbons generation, expulsion, and migration processes [25].

At the studied outcrop, we carried out the sampling of 20 representative hand samples along five key lithological horizons in the outcrop, named XP1, XP2, XG, AV, and ARE [see Fig. 2(b)], being four samples from each horizon. In general,

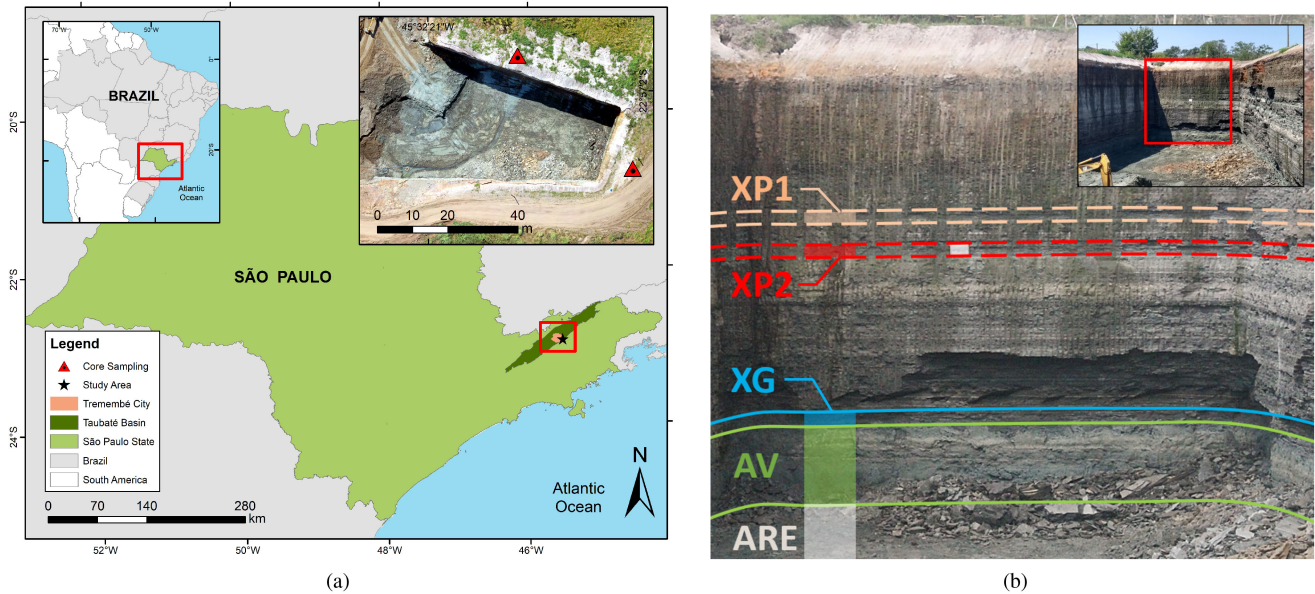


Fig. 2. Study area. (a) Location map and aerial image with drill-core location. (b) Outcrop sampled horizons.

the samples from the XP1, XP2, and XG horizons correspond to shales, the AV horizon is predominantly formed by clays with the presence of smectite (the minable portion of the mine), and the ARE horizon has the occurrence of clay and fine sand.

Among the collected samples, we selected five of them for laboratory analysis (one referent to each horizon), whereas the other 15 samples were reserved for spectral data collection. Moreover, we noted that samples from XP1 horizon were not homogeneous because they had differences in color between the two faces, indicating that they could have different mineralogy and organic content. Therefore, we choose to divide these samples into two aliquots (XP1a and XP1b) in order to confirm if this difference would be reflected in the collected data (geochemical and spectral).

Two drill cores were also collected near the mine area where the samples were collected [see Fig. 2(a)], named SF-01 and SF-02. The purpose of drilling for core extraction was the continuous and total sampling of rocks outcropped in the mine pit. Therefore, it would be possible to evaluate the characteristics of the rocks in all horizons present in the outcrop and not just those that had hand samples collected.

Although the drilling achieved 30 m, its recovery was very low, on average 15% for SF-01 and even lower for SF-02 with just 3%. Only the interval between 15 and 20 m of the SF-01 core had a better rock recovery with few missing intervals. However, even with the low recovery, it was possible to extract at least one sample every 1 m of drilling. As well as the outcrop samples, the SF-01 and SF-02 cores were also sent for laboratory and spectral analyses.

We included the images of the hand samples collected from the outcrop horizons and of the two recovered drill cores in the Supplemental Materials of this article.

B. Data Collection and Preprocessing

As the proposed method is based on supervised ML, it is necessary to collect a labeled dataset for models training and evaluation [26]. Therefore, after the sampling in an outcrop containing rocks with a high potential source of hydrocarbons (hand samples and drill cores, step A in Fig. 1), we performed laboratory analysis for geochemical and hyperspectral data collection (step B in Fig. 1).

The geochemical characterization of the rocks aimed to obtain data that describe the quantity and quality of organic matter from the TOC analysis and Rock-Eval pyrolysis. TOC contents were determined by combustion in a Leco SC-144DR carbon analyzer. The pyrolysis analysis was performed with a Rock-Eval 6 instrument, and from it was obtained the data of free or adsorbed hydrocarbons from the sample (S1 peak), hydrocarbons and CO₂ produced from kerogen cracking (S2 and S3 peak), and the temperature at the maximum generation rate of the S2 peak (Tmax). The hydrogen index (HI) and oxygen index (OI) were computed by the division of S2 and S3 by TOC, respectively. Those indices allow kerogen type determination through van Krevelen-type diagram [6].

To facilitate the geochemical analysis step, the samples collected on the same horizon of the Taubaté basin outcrop were considered identical (twins). Therefore, among those samples, one of them was selected for laboratory analysis (as a reference sample for its horizon), whereas the others were reserved for spectral data collection and ML experimentation.

In relation to the drill cores, each recovered fragment was considered a new sample. In the depth of 15 to 20 m from the SF-01, the collected fragments samples had an interval of 2 to 13 cm between them. This cores sampling resulted in 84 samples, 62 from the SF-01 core and 22 from SF-02 core. As it was not

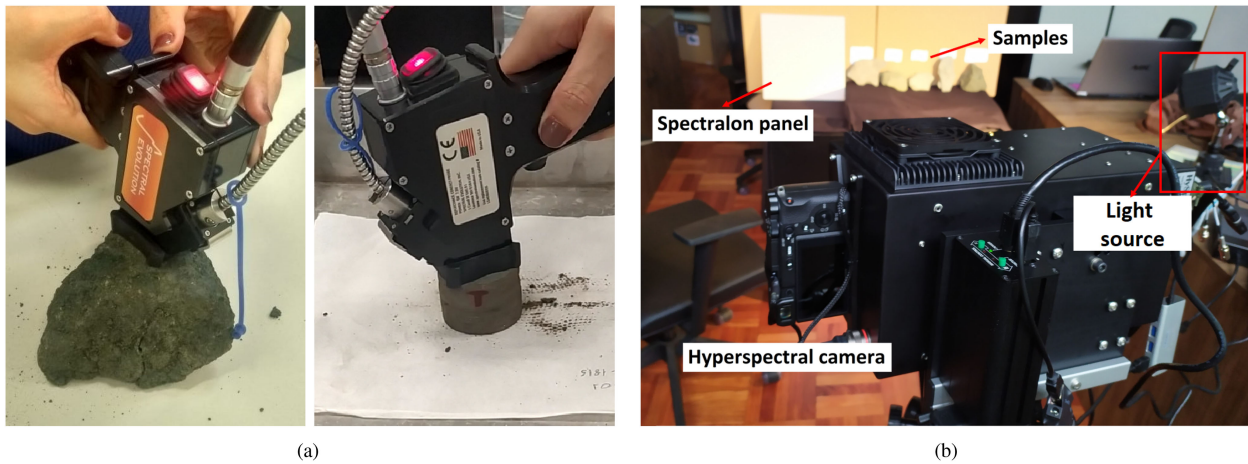


Fig. 3. Hyperspectral data collection. (a) Spectroradiometer Spectral Evolution. (b) Hyperspectral camera Mjolnir.

possible to obtain twin samples for the cores (one for destructive geochemical analysis and another for hyperspectral analysis), we first collected their spectral response and then sent them for laboratory analysis.

The hyperspectral data from the hand samples were collected using two distinct types of equipment, a nonimaging spectroradiometer and an imaging hyperspectral camera. Data from the drill core were measured only with the spectroradiometer for logistical reasons.

The nonimaging sensor used was the spectroradiometer Spectral Evolution, model SR-3500. This equipment acquired data in 1018 bands in a spectral range from 350 to 2500 nm, with spectral resolution between 3 and 7 nm. The methodology adopted was the spectral measurement of absolute reflectance using a contact probe [see Fig. 3(a)]. Therefore, before each new reading of a sample, the reference panel (Spectralon) was measured for reflectance correction.

Pointwise measurements were performed on the Taubaté Basin's samples and drill cores. We chose to perform more than one measurement on each sample, with the number of readings proportional to its size and availability of smooth surfaces on both sides. This process guaranteed that most of the sample was represented in the data collected.

The hyperspectral curves obtained through the spectroradiometer had some noise, mainly at the beginning and end of the reading range. Therefore, the data went through spectral smoothing using a moving average filter with a spectral window size equal to three bands. This filter produces a smoothed reflectance sr that is the mean value of the raw reflectance collected by the sensor in neighbor bands $sr_i = \frac{r_{i-1} + r_i + r_{i+1}}{3}$.

Samples imaging was carried out with the HySpex Mjolnir S-620 hyperspectral sensor. The experiment was set up in the laboratory with a 50 W tungsten halogen lamp (ILM-550, Spectral Evolution accessory) for artificial illumination at a distance of approximately 2 m between the sensor and the samples [see Fig. 3(b)]. The sensor's specification suggests a minimum distance of 20 m from the target. If used for smaller distances (as used in this work), the images may present a blurred aspect. As a

smaller distance is required to properly capture the hand samples given their size, in this study, we established a more controlled image acquisition setup in the laboratory and used an artificial light source. This greater control allowed us to compare more confidently the spectral signatures extracted from the images with those collected with a spectroradiometer and, then, confirm their usefulness for application in ML models.

After this disclaimer, the first step of image preprocessing was its correction for reflectance values, given that the image originally represents radiance values. The method used was the empirical line [27], in which the reference values for known targets in the image were used and a linear regression between radiance (raw data) and reflectance (targets) was applied to correct the image. In this study, the Spectralon panel was used as reference.

The second step was the spectral smoothing to reduce the noise present in the spectral curves of each pixel. This process was carried out with the application of the Savitsky–Golay filter, widely used in spectroscopy data. This filter consists of smoothing the spectral band based on a polynomial adjustment using the least squares method, considering a subset of data located in the filtering window, whose center is the data/band to be smoothed [28]. In this case, a filter with a window size three and a second-degree polynomial was considered to smooth the data with minimal loss of spectral information.

All of the steps of hyperspectral data preprocessing (round C in Fig. 1) were performed in Python environment (version 3.7) or using the software Envi (version 5.5).

C. Kerogen Type Classification

It is a common assumption in ML algorithms that training, validation, and testing data are independent and identically distributed. This assumption, although valid and necessary for many theoretical results, is seldom useful in practice [29]. In this study, two experiments are performed in the context of kerogen type classification using hyperspectral data: the first one explores how the trained classifiers are able to generalize

between distinct targets and the second one explores how a change in the acquisition sensor influences the classification results.

In the first experiment, SF-01 core data were used in training and validation, whereas the second core, SF-02, was reserved for testing. The second experiment is an extended version of the one published in [30] and includes the use of hyperspectral data collected from hand samples from the outcrop with spectroradiometer data used for training and validation of the models and the hyperspectral images as a test set.

The ML methods adopted for kerogen type classification in this study were: LR, KNN, RF, SVM, and MLP. Our selection of learners spans many distinct subareas within ML. We tested traditional statistical methods (LR), instance-based methods (KNN), theoretically optimal methods (SVM), ensemble methods (RF), and connectionist methods with neural networks (MLP).

The main steps consisted of: feature selection; dataset split for training and validation (cross validation); model training for tuning of hyperparameters (see Fig. 1-F); training and validation of the adjusted models (Fig. 1-E and G); and, finally, the evaluation of the performance for each classification model applied to the test dataset (Fig. 1-H).

The first step for the ML experiments was the selection of features for the models, which was performed by using domain knowledge. For this, we use the following criteria:

- 1) Reduce the analysis range to 1000–2500 nm, so both the spectroradiometer and the spectral camera have features in the same range.
- 2) Select bands that represent the absorption features reported in the literature about the spectral behavior characteristic of source rocks (clays, carbonates, and organic fraction) [7] that most likely could contribute with the classification.
- 3) Include the median and standard deviation of the selected features according to the sample to make the classifier more robust to noise in the data.

Knowing that Mjølner is able to acquire hyperspectral data in a shorter range of the spectra, the first criterion in the list above guarantees that we would be able to collect the same features using both sensors. The second criterion allowed us to use knowledge about the geochemical processes to select only a subset of the available features that most likely could contribute to the classification. The third and final criterion introduces two features that were engineered to make the classifier more robust to noise in the data.

To estimate the error of the trained classifiers, we adopted k -fold cross validation. In this approach, the dataset is divided into k disjoint subsets (folds) and then k runs are performed separating one of the folds for validation at each time.

Our dataset has characteristics that demand extra caution to perform the data sampling required for cross validation: there are several hyperspectral readings for each of the samples, the core data are ordered with semantic meaning (by depth) and the target value is not balanced. Two caution steps are performed in the data sampling to guarantee a proper error estimation under this circumstance: 1) stratified grouped sampling needs to be

TABLE I
TUNED HYPERPARAMETERS IN ML EXPERIMENTS

Classifier	Hyperparameters	Values
LR	C	0 to 100 (uniform, float)
	l1_ratio	0 to 1.0 (uniform, float)
KNN	n_neighbors	1 to 9 (uniform, int)
	metric	manhattan, euclidean, chebyshev, angular
	weights	uniform, distance
RF	n_estimators	0 to 500 (uniform, int)
	criterion	gini, entropy
	min_samples_leaf	{1, 2, 3}
	ccp_alpha	0 to 100 (uniform, int)
SVM	C	0 to 100 (uniform, int)
	Kernel	linear, rbf, poly
	Degree	{2, 3, 4, 5, 6}
MLP	Gamma	{auto, scale}
	hidden_layer_sizes	1, 2 or 3 layers, {1 to 200} neurons
	alpha	0 to 0.1 (uniform, float)
	learning_rate	constant, adaptive
	learning_rate_init	0 to 0.01 (uniform, float)
batch_size	1 to 10 (uniform, int)	

performed to guarantee that the same sample does not appear represented in training and in validation data (as this would be data leakage) while also keeping the target value distribution roughly the same among folds; 2) data need to be shuffled before the folds are created, otherwise we would have a cross-validation scheme where each of the folds represents specific depths in the core data, which in turn would make the cross-validation error estimation more pessimistic.

To tune the hyperparameters of the classifiers, we adopted the random search strategy with tenfold stratified cross validation. Random search was chosen instead of traditional grid search because it explores the search space randomly, being more robust in cases where some hyperparameters do not affect the results while also allowing us to define a limit to the computational budget [31]. The method is executed in the following manner: for each of the hyperparameters in the search space, we defined a sampling distribution, as described in Table I. In each of the runs, the hyperparameter value was sampled from the defined distribution. We also defined the maximum compute budget to be 1000 executions. In cases such as the MLP, we stopped the experiment at 1000 runs, whereas in the case of learners with less than 1000 combinations, such as the KNN, the search exhausted all the possibilities (similar to a grid search setup). The best set of hyperparameters for each learner was the set that resulted in the highest f1-score averaged on the three classes.

To evaluate how the models classified each kerogen type and their general performance, the following metrics were computed in the validation set: overall accuracy (Acc), Kappa coefficient, precision, recall, and f1-score. The definition of these metrics is found as follows:

$$\text{Acc} = \frac{1}{N} \sum_{i=1}^N I(y^{(i)} = \hat{y}^{(i)}) \quad (1)$$

$$\text{Kappa} = \frac{(p_o - p_e)}{(1 - p_e)} \quad (2)$$

$$\text{Precision} = \langle P(y_c, \hat{y}_c) | c \in C \rangle \quad (3)$$

$$P(A, B) = \frac{|A \cap B|}{|A|} \quad (4)$$

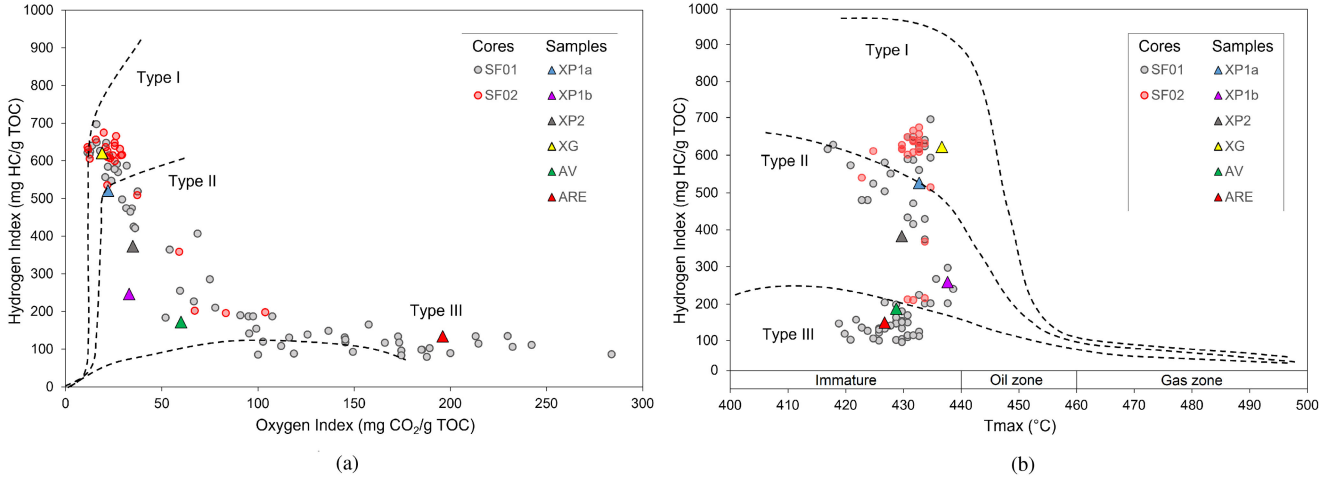


Fig. 4. Geochemical results from the six hand samples (colored triangles) and two cores collected (gray and red circles). (a) Modified Van Krevelen diagram to indicate the kerogen types. (b) HI versus Tmax graph to indicate source rock thermal maturity.

$$\text{Recall} = \langle R(y_c, \hat{y}_c) | c \in C \rangle \quad (5)$$

$$R(A, B) = \frac{|A \cap B|}{|B|} \quad (6)$$

$$F_{\beta_score} = \langle F_{\beta}(y_c, \hat{y}_c) | c \in C \rangle \quad (7)$$

$$F_{\beta}(A, B) = (1 + \beta^2) \frac{P(A, B) \times R(A, B)}{\beta^2 P(A, B) + R(A, B)} \quad (8)$$

where $C = \{I, II, III\}$, y_c and \hat{y}_c are the set of observations and predictions from class c , respectively, and β is the parameter that controls the F_{β_score} , being defined $\beta = 1$ in this study, so we track the harmonic mean between Precision and Recall.

Finally, the five trained models (one for each classifier) were applied to the test set and the same evaluation metrics were computed. In the case of the first experiment, the data came from another core and in the case of the second experiment, the data came from hyperspectral images from the Taubaté Basin's samples.

For model testing in the hyperspectral images, the bands related to the wavelengths chosen as features for the models were extracted from each pixel of the images and then used for inference of a kerogen type class for each corresponding pixel. As a result of this process, classified images according to the kerogen type were obtained for each sample.

The implementation of this ML step was performed in a Python environment (version 3.7) using the scikit-learn library (version 1.0.1) and tracked using the MLOps platform Weights and Biases. As support for the application of the models in hyperspectral images and their subsequent visualization, it was used the software ArcGIS (version 10.6.1) and ENVI (version 5.5).

III. RESULTS AND DISCUSSION

A. Geochemical Data

The geochemical analysis results showed different rocks' characteristics regarding the presence of organic matter. In hand

samples collected at the outcrop, TOC values ranged from 0.3% and 0.5% in claystones (ARE and AV samples, respectively) to 4.0% and 4.3% in shales (XP1a and XG samples, respectively). Generally, for thermally immature source rocks (such as those from Fm. Tremembé), TOC values starting at 4% are considered excellent [3]. Due to the more significant amount of data collected, the range of core results is greater with high TOC contents (greater than 10%) found in papyraceous shales at the beginning of the profile (up to 7 m).

As a result of the relationship between the geochemical analyses of TOC and Rock-Eval pyrolysis, the HI and OI indices were computed and plotted on the Modified Van Krevelen Diagram [6] for kerogen type classification [see Fig. 4(a)]. We noticed that the hand samples collected in the outcrop are distributed in one sample as Type I (sample XG), three as Type II (XP1a, XP1b, and XP2), and two as Type III (AV and ARE). The data from the SF-01 core are well distributed in the three classes, while the SF-02 core presented only Type I and II samples. The samples of this study present interesting variations in facies (shales, siltstones, and claystones) in the preservation state of the organic matter, probably in organic matter type and, consequently, in the kerogen.

Moreover, in Fig. 4(b), we present a diagram between HI and temperature at maximum hydrocarbon generation rate (Tmax) [6]. This graph indicates the low degree of thermal maturation of the analyzed samples, which are considered to potential source rocks.

To analyze the kerogen type behavior in the outcrop in a sequential way (according to the depth), we present Fig. 5.

In Fig. 5, the regions of the geochemical profile of the SF-01 core corresponding to the sampled horizons XP1, XP2, XG, AV, and ARE are highlighted. One aspect to be highlighted is that, although the AV and ARE horizons are predominantly composed of Type III kerogen, they have intercalations and mixtures of Type II kerogen resulting from the entry of distinct organic matter into the lake during its formation.

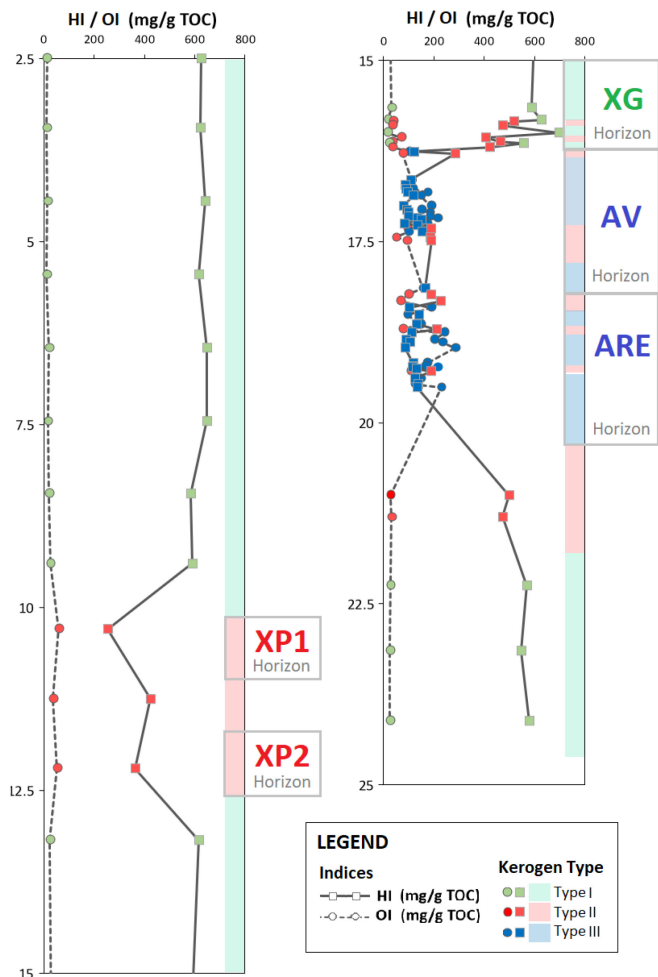


Fig. 5. Geochemical profile of the SF-01 core with OH and HI plotted. The horizons where samples were collected are demarcated in the image. The colors indicate the kerogen type.

B. Hyperspectral Data

The process of hyperspectral collection resulted in a spectral library with 163 reflectance curves from cores and 91 from hand samples. The mean values of some spectral curves from SF-01 core and from hand samples are presented in Fig. 6. Full results can be viewed in the Supplemental Materials added in the Appendix.

In order to use hyperspectral data in the characterization of source rocks, we need to know the spectral signature of the different minerals and compounds that constitute these rocks, mainly through reference libraries. However, since these rocks are sedimentary (constituted by a large mixture of minerals), their reflectance spectrum is composed of the superposition of all the spectra of their components [12]. Thus, to establish a relationship between the spectral responses of these potential source rocks with the parameters of interest, this mixture of spectra must also be considered.

In Fig. 6, we note that although the mean curves for each sample show their differences, it is possible to see the presence of absorption features in 1400, 1900, and 2200 nm in most samples. The presence of these features occurs throughout the

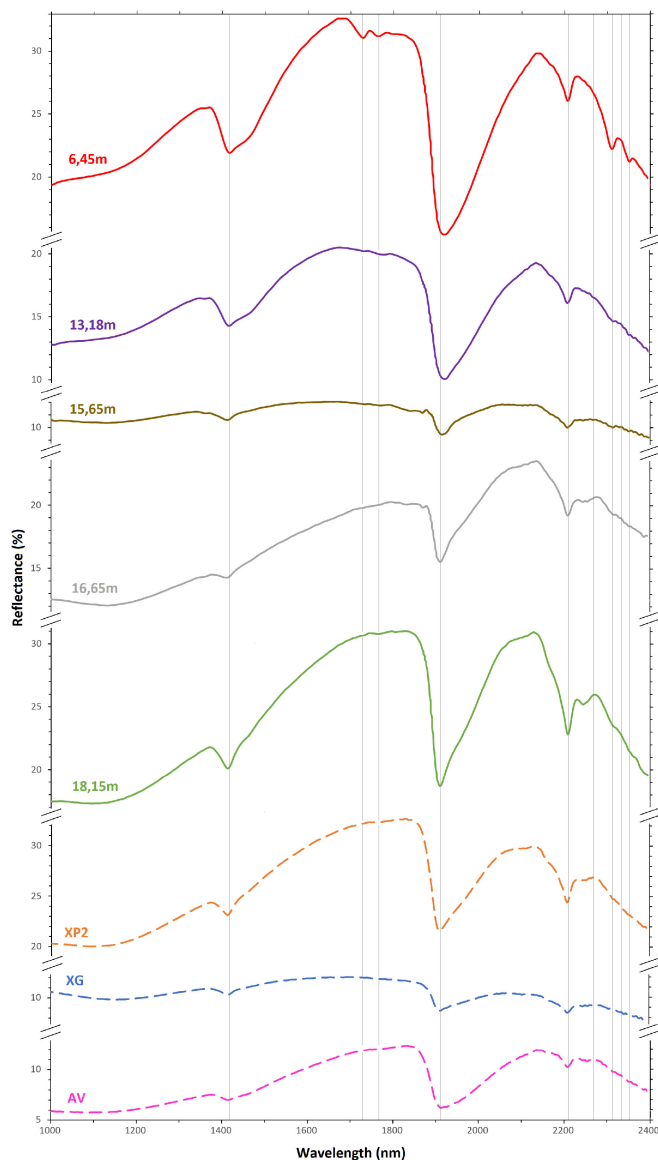


Fig. 6. Some mean spectral reflectance curves collected from the core (solid line) and from the hand samples (dotted line). Full results can be viewed in the Supplemental Materials.

entire hyperspectral dataset, collected both in the hand samples and drill cores, and has a relation with molecular water and hydroxyl ion, common in clay minerals [7], [32], [33]. As the outcrop is from a lacustrine environment, all layers present a clay composition.

Among the spectral library collected, only a few samples of the cores showed the absorption bands at 1700 and 2300 nm, indicated in the literature as characteristics of the presence of bitumen in rocks [7], [13]. These features are observed in the spectral signature at 6.45 m of the core's depth (see Fig. 6), whose laboratory analysis indicated a TOC concentration of 8.1% and Type I kerogen. On the other hand, the spectral responses at 13.18 m core's depth (6.2%) and XG hand sample (4.3%), which showed a lower TOC concentration in the geochemical analyses, did not show the features mentioned.

TABLE II
 SPECTRAL ANGLE SIMILARITY—HAND SAMPLES

	Type I	Type II	Type III
Type I	1	0.916	0.798
Type II	0.916	1	0.866
Type III	0.798	0.866	1

 TABLE III
 SPECTRAL ANGLE SIMILARITY—SF-01 DRILL CORE

	Type I	Type II	Type III
Type I	1	0.934	0.857
Type II	0.934	1	0.912
Type III	0.857	0.912	1

In addition to the bands related to the organic content and clay minerals, in some spectral curves was also identified an absorption band at 2350 nm, which is a spectral feature of carbonate minerals. This feature may be related to the presence of ostracod, which are fossils of small crustaceans present in rocks whose characteristic is the presence of a calcified bivalve carapace [34]. The presence of these fossils was observed in the samples' visual analysis, agreeing with the spectral curves that presented the mentioned absorption band.

The similarity between the collected hyperspectral data was analyzed using the similarity function defined in (9), which is the same similarity that is used in the spectral angle mapper—SAM algorithm [35]. We compared the spectral libraries of this study with each other and confirmed what is observed in Fig. 6. Among the outcrop samples, the lowest correspondence index was greater than 0.8, indicating high similarity between the spectra, even though they may have differences in their main mineralogical composition or their amount and type of organic matter. As the core data were more heterogeneous, when we analyzed this spectral library, the minimum similarity values were up to 0.45 (between a papyraceous shale type I with 16% TOC and that of a type III claystone with 0.9% TOC).

$$SA(X^{(i)}, X^{(j)}) = \arccos \left(\frac{X^{(i)} \cdot X^{(j)}}{\|X^{(i)}\| \|X^{(j)}\|} \right). \quad (9)$$

To evaluate the spectral behavior of kerogen types, we computed the mean reflectance values and standard deviation among all samples of each class (see Supplemental Materials, Fig. 15). The similarity between the kerogen type spectra was also analyzed using SAM and the results are presented in Tables II (for outcrop samples) and III (SF-01 core). Likewise, the comparison of similarities between the samples and cores spectra, here we also observed the spectral proximity between the kerogen types with the lowest correspondence between the classes of Types I and III (0.798 for samples and 0.8 for core), mainly due to the impact that the change of facies and the presence of organic matter cause in the spectrum.

This high spectral similarity between the samples and, consequently, between the different kerogen types is indicative of the complexity of the problem approached in this article and corroborates the need to use robust techniques for its classification.

At last, we present the hyperspectral images collected with the Mjolnir S-620 sensor and preprocessed according to the

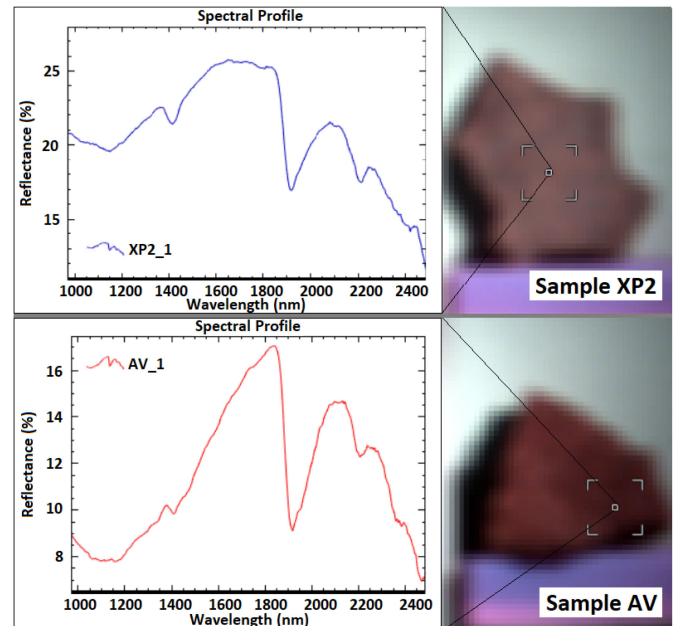


Fig. 7. Hyperspectral images of XP2 and AV samples and an example of its reflectance curves.

 TABLE IV
 BANDS ADOPTED AS FEATURES FOR ML MODELS

Wav. (nm)	Name	Description
1412	C11	Clay absorption band
1729	B1	Bitumen absorption band
1750	-	Between bands B1 and B2
1765	B2	Bitumen absorption band
1909	C12	Clay absorption band
2207	C13	Clay absorption band
2271	-	Between bands C13 and B3
2310	B3	Bitumen absorption band
2333	-	Between bands B3 and Ca
2350	Ca	Carbonate absorption band

methodology previously explained. Fig. 7 shows the hyperspectral image from two samples and an example of the reflectance curve extracted from one pixel of each image. Experiments showed that for this considered acquisition setup, the effective pixel is on the order of 1 cm.

The blurred effect on the images is due to the small distance used in this study between sensor and target (2 m instead of the 20 m minimum distance indicated by HySpex), which can hinder the clear visualization of samples' edges and small features on their surface. However, we performed validation on Mjolnir's reflectance data by inspecting the spectral curves extracted in image pixels and comparing them with the data obtained by the spectroradiometer. Therefore, when observing Figs. 6 and 7, we noted that the spectral signatures of the samples acquired in the XP2 and AV horizons are similar.

After analyzing the spectral signature of all the outcrop samples (hand samples and drill cores data), we selected ten features (bands) that could be important for the classification of kerogen type in source rocks. The features used as input of the models are presented in the Table IV and in Fig. 6.

TABLE V
SELECTED HYPERPARAMETERS IN EXPERIMENTS 1 AND 2

Classifier	Hyperparameters	Exp. 1	Exp. 2
LR	C	34.858	0.7249
	ll_ratio	0.1253	0.705
KNN	n_neighbors	1	1
	metric	angular	chebyshev
	weights	uniform	distance
RF	n_estimators	285	732
	criterion	gini	gini
	min_samples_leaf	1	1
	ccp_alpha	0	0
SVM	C	41.728	78.091
	Kernel	linear	linear
	Degree	-	-
	Gamma	scale	scale
MLP	hidden_layer_sizes	(104, 9)	(139, 104, 100)
	alpha	0.07818	0.07274
	learning_rate	constant	constant
	learning_rate_init	0.003955	0.000478
	batch_size	6	6

TABLE VI
RESULTS ML EXP. 1 VALIDATION

Classifier	Acc	Kappa		Precision	Recall	f1-score
LR	0.81	0.70	I	0.87	0.83	0.85
			II	0.80	0.78	0.79
			III	0.79	0.84	0.82
KNN	0.81	0.70	I	0.88	0.88	0.88
			II	0.80	0.78	0.79
			III	0.79	0.81	0.80
RF	0.71	0.54	I	0.79	0.63	0.70
			II	0.69	0.74	0.71
			III	0.71	0.73	0.72
SVM	0.84	0.75	I	0.92	0.92	0.92
			II	0.82	0.82	0.82
			III	0.81	0.81	0.81
MLP	0.83	0.73	I	0.91	0.88	0.89
			II	0.80	0.82	0.81
			III	0.81	0.81	0.81

By proposing as the model's features not only the bands directly related to the organic matter (B1, B2, and B3) but also those characteristics of inorganic minerals constituting these rocks (Cl1, Cl2, Cl3, and Ca); instead of trying to isolate organic matter and ignoring the influence of these minerals in the spectrum, we sought to take advantage of possible relationships they might have with our target variable (kerogen type).

C. Experiment 1: Core

In this first experiment, data from drill cores were explored with ML algorithms to classify potential source rocks based on its kerogen type. After numerous runs performed, the set of hyperparameters selected for each algorithm is presented in Table V.

In Table VI, we present the results of the model validation step. With the exception of RF, all other classifiers had Acc and Kappa greater than 0.8 and 0.7, respectively. We emphasize SVM and MLP, which presented the highest values of all metrics computed. Application of the SVM algorithm was also proposed

TABLE VII
RESULTS ML EXP. 1 TEST

Classifier	Acc	Kappa		Precision	Recall	f1-score
LR	0.85	0.65	I	1.00	0.80	0.89
			II	0.60	1.00	0.75
			III	-	-	-
KNN	0.77	0.53	I	1.00	0.70	0.82
			II	0.57	1.00	0.73
			III	0.00	-	-
RF	0.42	0.13	I	1.00	0.25	0.40
			II	0.29	1.00	0.44
			III	-	-	-
SVM	0.85	0.65	I	1.00	0.80	0.89
			II	0.60	1.00	0.75
			III	-	-	-
MLP	0.71	0.43	I	1.00	0.63	0.77
			II	0.44	1.00	0.62
			III	-	-	-

in [36] for a robust oil estimation method in oil shale samples showing promising results.

Focusing on the results by class, we noticed that the best performance of most models was in the classification of Type I. This is an encouraging result because this type is the one with the greatest interest in hydrocarbon exploration, indicating the rocks with a high generation potential [3], [6]. In addition, another important aspect is that when analyzing the models' confusion matrices,¹ LR, KNN, SVM, and MLP did not present confusion in the classification between Types I and III (opposite classes in terms of generating potential).

To evaluate the models ability to generalize what had been learned to targets different from those used in their construction (data from SF-01 as training and validation), we submitted them to a new dataset from another core collected in the same study area (SF-02). The computed metrics of this test are presented in Table VII.

With the application of a new dataset to the trained models, some degradation in their performance is expected, but it should not be significant to the point of generating distrust of overfitting in the learner. Although the SF-02 core dataset does not have samples of Type III kerogen, the results presented in the table show that LR and SVM performed well, both with Acc greater than 0.8 and Kappa greater than 0.6. Moreover, while MLP had a high result in the validation step, when applied to the new dataset its performance decreased a lot (Kappa from 0.73 to 0.43). Cases such as the MLP, with a considerable decrease in the metrics from validation to testing indicate that there is some overfitting. This happens because, although spectral data from distinct cores are similar they are not the same (see Supplemental Materials, Figs. 21 and 22 for a visualization of those differences); on the other hand, the SVM appears to be able to generalize to distinct cores.

Regarding the Type I and II classes, we highlight that the five models had precision = 1 for Type I kerogen (i.e., every prediction in this class was assertive) and recall = 1 for Type II (i.e., samples from this group were 100% classified correctly).

¹The confusion matrices can be checked in the Appendix (Supplementary Material).

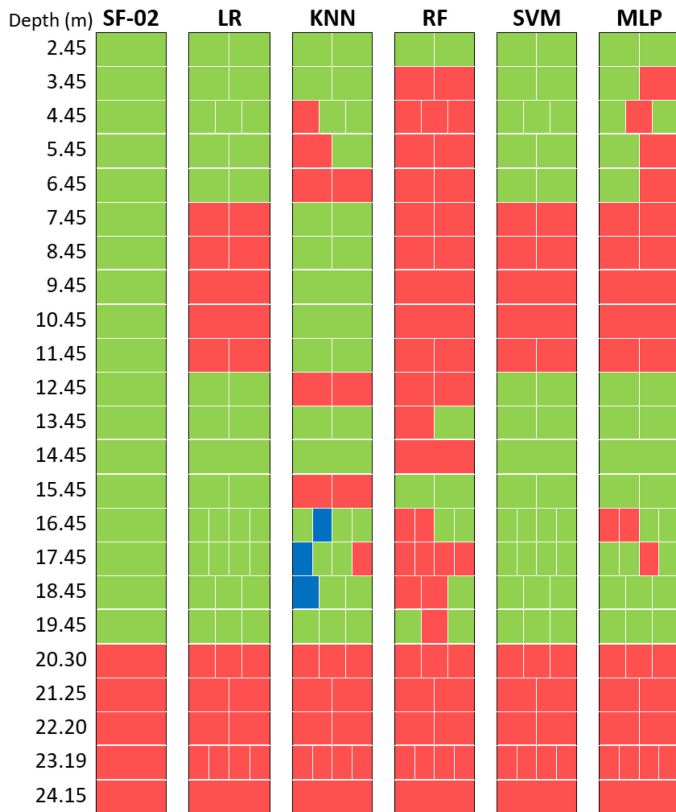


Fig. 8. Representative image of the true classification of the SF-02 core and the results of each model (unscaled image). The subdivisions in each line/depth indicate spectral measurements performed on the same drill core fragment (at the top, bottom, and side positions, for example).

To instance the results mentioned and listed in Table VII, we present Fig. 8. This image illustrates the comparison between classes predicted by each algorithm from the spectral signatures of the SF-02 fragments in different depths and its real kerogen classification.

Analyzing Fig. 8, we highlight two interesting observations. The first is that KNN, RF, and MLP presented disagreeing classifications for spectral measurements of the same sample/depth. For example, the MLP model predicted different classes for the two spectral curves from the 3.45 m fragment (3.45_1 as Type I and 3.45_2 as Type II). This indicates that the threshold between classes of the mentioned models may not have been adjusted enough to hyperspectral data from the same samples (and, consequently, very similar).

Also, four learners, among them LR and SVM which had the best performance, missed the prediction in the same samples (from depths 7.45 to 11.45 m). In an attempt to understand this behavior, we first analyzed the geochemical data that determined these samples as Type I kerogen, but all of them had HI greater than 600 mg HC/g TOC, and therefore, they are not on the threshold between the two groups. Eliminating the first possibility, we looked at the models input data: the spectral signatures. In Fig. 9, we present some curves of the SF-02 core plotted in the wavelength range from 1200 to 2400 nm. Observing the five curves of interest (represented by the yellow lines), it is

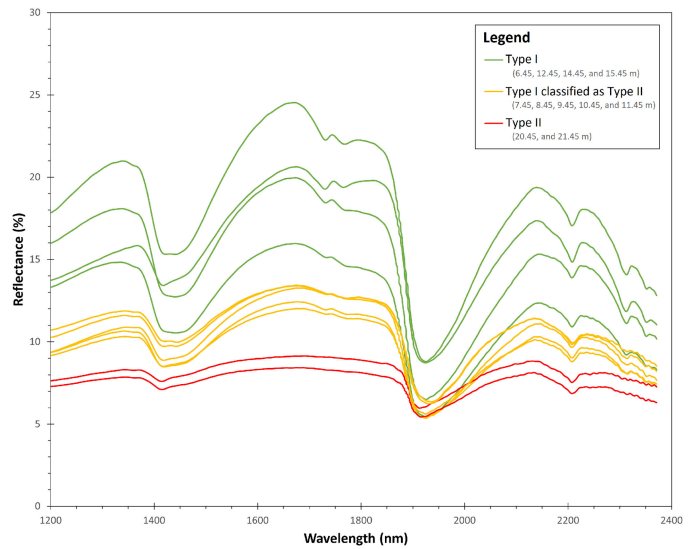


Fig. 9. Comparison between some spectral curves of the SF-02 core. The colors represent: in green are some Type I samples whose LR and SVM models correctly classified; in yellow are samples from depths of 7.45 to 11.45 m that are Type I and were classified as Type II; in red some samples correctly classified as Type II.

noticed that they are visibly different from the others of the same class (green lines), mainly in the reflectance intensity and in the behavior of the two regions of absorption bands characteristic of organic matter at 1700 and 2300 nm. Indeed, the curves seem to have an intermediate behavior between those of Type I and II (red lines). Therefore, we assume that these misclassifications were motivated by the resemblance between the spectra of the 7.45 to 11.45 m samples with those of Type II (lack of 1700 and 2300 nm absorption bands, for example).

D. Experiment 2: Sample

The results of the previous experiment were the first step for us to state that kerogen type classification using hyperspectral data is possible. Here, we went further using the models trained with spectroradiometer to classify images collected by an imaging sensor, in this case Mjolnir Hyperspectral camera. Table I lists the hyperparameters selected for this second experiment.

In Table VIII, we present the results of the model validation step. SVM and MLP also had the best performances, such as in the first experiment, both with Acc of 0.94 and Kappa of 0.90. When we look at the metrics by class (and the confusion matrices in the Supplemental Materials), we notice that 100% of Type I and III samples were correctly classified (recall = 1) and all Type II and III prediction was correct (precision = 1).

Table IX presents the results of the Experiment 2 test, i.e., the application of trained models in the hyperspectral images of hand samples. Two highlights can be made when inspecting these results: MLP was kept as the best model; and the LR model, which had lagged in the validation, fit well with the images and ranked second among the classifiers in the test step. For us, this pattern that happened with LR result is not interesting because the validation result of this model was the best in the

TABLE VIII
RESULTS ML EXP. 2 VALIDATION

Classifier	Acc	Kappa		Precision	Recall	f1-score
LR	0.77	0.63	I	0.63	0.53	0.57
			II	0.70	0.85	0.77
			III	1.00	0.81	0.89
KNN	0.78	0.66	I	0.71	0.79	0.75
			II	0.80	0.80	0.80
			III	0.79	0.74	0.77
RF	0.89	0.83	I	0.71	0.79	0.75
			II	1.00	0.85	0.92
			III	0.89	1.00	0.94
SVM	0.93	0.90	I	0.76	1.00	0.86
			II	1.00	0.85	0.92
			III	1.00	1.00	1.00
MLP	0.93	0.90	I	0.76	1.00	0.86
			II	1.00	0.85	0.92
			III	1.00	1.00	1.00

TABLE IX
RESULTS ML EXP. 2 TEST

Classifier	Acc	Kappa		Precision	Recall	f1-score
LR	0.84	0.76	I	1.00	1.00	1.00
			II	0.70	1.00	0.82
			III	1.00	0.47	0.64
KNN	0.79	0.68	I	0.85	0.91	0.88
			II	0.70	0.99	0.82
			III	0.96	0.42	0.58
RF	0.60	0.38	I	0.78	0.59	0.67
			II	0.64	1.00	0.78
			III	0.19	0.12	0.15
SVM	0.79	0.68	I	1.00	0.99	0.99
			II	0.63	1.00	0.78
			III	1.0	0.30	0.47
MLP	0.86	0.79	I	1.00	1.00	1.00
			II	0.72	1.00	0.84
			III	1.00	0.54	0.70

hyperparameters random search, and then, this superior result in the test may indicate some poor fit to the data that could lead to poor generalization. Differences in the distribution of the bands can be visualized in the Supplemental Materials (Figs. 19 and 20).

When we applied the models to the hyperspectral images, some differences regarding the precision and recall metrics between the validation (see Table VIII) and test (see Table IX) were identified. To help these results discussion, it is important to analyze, together with the tables, the classified images of each sample. We show in Fig. 10 the classified images of ten samples for the five models and their ground truth obtained by laboratory analysis. We highlighted two results: the worst (RF with Acc = 0.60; Kappa = 0.38) and the best (MLP with Acc = 0.86; Kappa = 0.79).

The first issue about the images presented in Fig. 10 is the uniformity of the classification inside the same sample. Preliminary results published in [30] showed very noisy classified images, which caused some confusion in its interpretations. Here, we decided to perform feature engineering to extract the mean and std values for each band selected as input into models to try attenuate this problem. Therefore, when including as features the median and standard deviation of these bands for each sample, we made the classifier more robust to noise in the data.

The same behavior observed in the first experiment also occurred in this one: the excellent adjustment of the learners to classify the Type I kerogen samples and the good ability to distinguish the Type I and Type III classes. First, as Type I is the best for hydrocarbons generation, our models working well for this class is a good indication of the functioning and applicability of the classifier. It is even more encouraging if we also bring out the aspect of errors (misclassification) occurring only between adjacent classes, i.e., Types I and II, and Type II and III. These results indicate that the classifiers correctly learned the relationship between the hyperspectral data and the hydrocarbon generation potential of the samples, reducing the possibility of the high metrics in the validation stage, were due to some spurious correlation.

The low recall values for class Type III listed in Table IX are evident when looking at the classified images in Fig. 10. It appears that the ARE samples are the hardest ones to classify; only kNN was able to correctly classify one of ARE samples, the other models misclassified them as Type II. The MLP results (best model) are slightly different. The spectral signature of one face of each ARE sample made the model classify it as two kerogen types spatially distributed, in some portions as Type II and others as Type III.

In the evaluation of the classified images, we expect to see only one class for each sample. Although there is spatial variation in the content of organic matter within the sample, this should not be significant to the point of changing its kerogen type. However, when analyzing the result of the ARE sample classified by the MLP algorithm (see Fig. 10), the hypothesis of mixing two classes in the same sample was considered. This fact may be related to episodes of sedimentary transition of the lake in which zones of mixing occur close to the events of sedimentation transition, thus occurring the mixing of different organic materials, resulting in kerogen of different compositions. In the SF-01 drill core profile (see Fig. 5), it is possible to observe the mixture of kerogen types present in the respective ARE horizon.

These reported inconsistencies in relation to the different methods for kerogen classification (geochemical and hyperspectral) occur due to the analyses nature. The laboratory data are obtained by a punctual and uniform sampling, where a volume of the sample is powdered and homogenized to obtain the result, which is extrapolated for all the sample. In the other hand, the hyperspectral image analyzes the surface of a target, being considered as a broader sampling, given that each point (in this case, pixel) has a respective value with no need for data extrapolation.

Considering the above-mentioned results and discussion, the use of hyperspectral images in applications like this experiment clearly further expands the potential of the nondestructive classification method proposed here. Besides, inferring the kerogen type in the sample (approach with a nonimaging sensor as a spectroradiometer), this information can also be analyzed spatially. In addition to the application of hyperspectral images in the sample scale, helping with a fast characterization of them without the need for laboratory analysis, this can be even more useful in the outcrop scale, being extremely important and useful

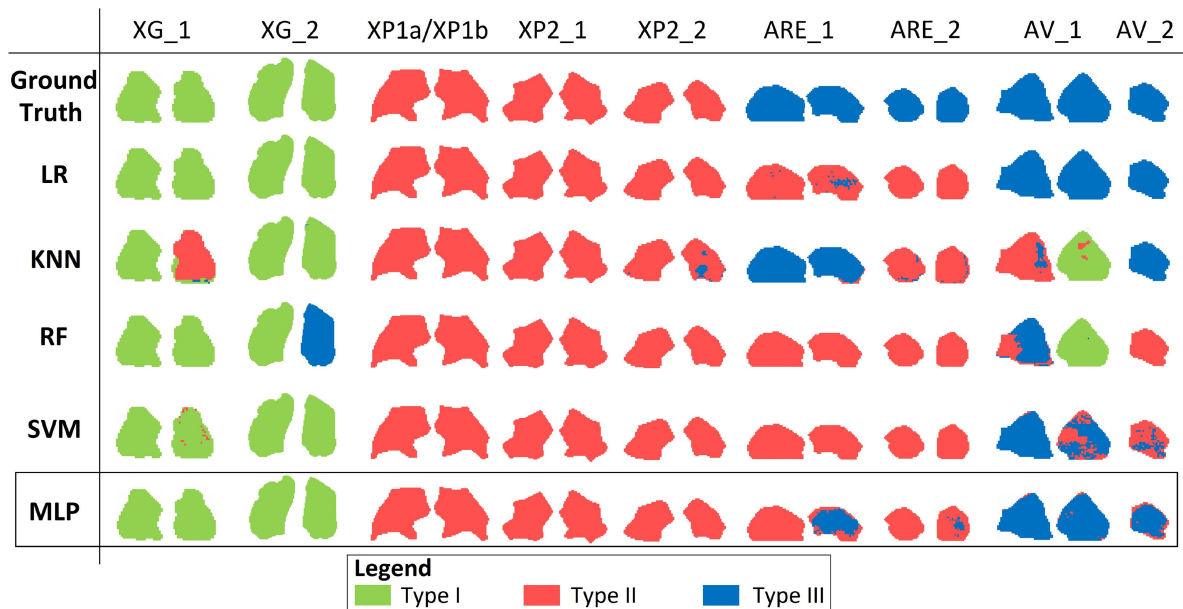


Fig. 10. Classified images for the five models for Taubaté samples and their ground truth obtained by laboratory analysis. The highlighted result (MLP model) had the best performance.

for understanding the kerogen type's spatial behavior in the outcrop and also for sampling orientation [8].

E. Remarks About the Experiments

Differences in the results between learners and experiments were expected. Each learning algorithm used in the experiments has inductive biases, that is, it prioritizes some functions from the hypothesis space from others when trying to approximate the relation between the reflectance spectra and the kerogen type.

Considering Tables VII and IX, there is discrepancy between the results of the same learner across experiments (e.g., MLP's accuracy in testing is 0.71 and 0.86 for Experiments 1 and 2, respectively). This is due to the fact that each of the experiments explores how the models generalize in distinct situations. In Table VII, we present the results obtained using the trained models in a new/unseen core, whereas Table IX represents the results obtained in the same samples but using a distinct hyperspectral sensor. This difference in experiment setting shows us that the same learner (e.g., the MLP) behaves differently when used in a different setup than it was trained on; from our experiments, the generalization between sensors appears to be easier than generalization between targets. The LR learner appeared to be more robust to changes than the others; this was expected because LR is a high-bias low-variance learner.

The method presented in this study proved to be applicable for kerogen type classification in potential source rocks (immature rocks). The samples analyzed here represent a lacustrine depositional system with variations in the facies in the preservation state of the organic matter and, probably, in the type of organic matter. The effect that other types of lithologies would have on trained models was not evaluated here, and therefore, our models may be limited to formations with characteristics similar to the

study area. To apply the method presented in this study to other sedimentary basins, it is vital that the geochemical and hyperspectral dataset used in the models' training is representative of these new areas; otherwise, our results will not be replicable. Future studies will aim on expanding the dataset to account for more sedimentary basins and different degrees of thermal maturation.

IV. CONCLUSION

The kerogen type is closely related to the hydrocarbon generation potential. Therefore, defining it is a critical step in source rock characterization. Traditional methods to determine the kerogen type have significant drawbacks: they require specialized tools and personnel and are destructive. Our results are prospective and show that ML techniques applied to the hyperspectral data were useful to classify potential source rocks according to their kerogen type. The method presented in this study provides researchers and practitioners with an alternate procedure that alleviates the bottlenecks of traditional methods to determine kerogen type. It is fast, nondestructive, and, to the extent of our testing, appears to be robust in distinct hyperspectral sensors.

We performed two experiments in this study, both using spectral signatures collected with spectroradiometer in core samples and hand samples to train and validate several supervised ML models, most achieving accuracy above 0.8 in the validation step. We presented evidence that the models trained are robust, as they were able to generalize to datasets with different targets and to data collected by another hyperspectral sensor. Furthermore, we show the interesting ability of classifiers to correctly define Type I samples and mainly distinguish the kerogen type between Type I (high generation potential) and Type III (low potential) samples.

Rapid and synoptic techniques for the inference of geochemical characteristics of source rocks, as proposed in this work, have great potential to be translated into real-world applications. Some benefits of this approach are: to simplify sample screening, allow estimation of organic matter quality indirectly, minimize operation and time costs, and avoid possible errors caused by discrete and punctual sampling in heterogeneous geobodies.

Finally, it is not expected that the models trained in this study would be immediately able to generalize to rock formations with physicochemical characteristics not represented in the dataset. On the contrary, the model generalization error is often bounded by the quality and diversity of the sampling procedure. It is vital to have access to a representative dataset of the area studied before using this method.

ACKNOWLEDGMENT

The authors used Weights and Biases for experiment tracking and visualizations to develop insights for this article.

REFERENCES

- [1] B. P. Tissot and D. H. Welte, *Petroleum Formation and Occurrence*. Berlin, Germany: Springer Science & Business Media, 2013.
- [2] E. J. Milani, J. Brandão, P. Zalán, and L. Gamboa, "Petróleo na margem continental brasileira: Geologia, exploração, resultados e perspectivas," *Revista Brasileira de Geofísica*, vol. 18, pp. 352–396, 2000.
- [3] K. Peters, X. Xia, A. Pomerantz, and O. Mullins, "Geochemistry applied to evaluation of unconventional resources," in *Unconventional Oil and Gas Resources Handbook*. New York, NY, USA: Elsevier, 2016, pp. 71–126.
- [4] S. D. Killops and V. J. Killops, *Introduction to Organic Geochemistry*. Hoboken, NJ, USA: Blackwell, 2005.
- [5] N. M. Al-Areeq, "Petroleum source rocks characterization and hydrocarbon generation," in *Recent Insights in Petroleum Science and Engineering*. Rijeka, Croatia: IntechOpen, 2018.
- [6] J. Espitalié, M. Madec, B. Tissot, J. Mennig, and P. Leplat, "Source rock characterization method for petroleum exploration," in *Proc. Offshore Technol. Conf.*, 1977, Paper OTC-2935-MS.
- [7] M. Speta, B. Rivard, J. Feng, M. Lipsett, and M. Gingras, "Hyperspectral imaging for the determination of bitumen content in Athabasca oil sands core samples," *AAPG Bull.*, vol. 99, no. 7, pp. 1245–1259, 2015.
- [8] Y. Mehmani, A. K. Burnham, M. D. Vanden Berg, and H. A. Tchelepi, "Multiscale characterization of spatial heterogeneity of petroleum source rocks via near-infrared spectroscopy," in *Proc. Unconventional Resour. Technol. Conf.*, Austin, TX, USA, 2017, pp. 2539–2543.
- [9] I. C. C. Acosta, M. Khodadadzadeh, R. Tolosana-Delgado, and R. Gloaguen, "Drill-core hyperspectral and geochemical data integration in a superpixel-based machine learning framework," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 4214–4228, Jul. 2020. [Online]. Available: <https://ieeexplore.ieee.org/document/9146200>
- [10] S. Rajendran, "Mapping of neoproterozoic source rocks of the Huqf supergroup in the sultanate of Oman using remote sensing," *Ore Geol. Rev.*, vol. 78, pp. 281–299, 2016.
- [11] Y. Mehmani, A. K. Burnham, M. D. V. Berg, F. Gelin, and H. Tchelepi, "Quantification of kerogen content in organic-rich shales from optical photographs," *Fuel*, vol. 177, pp. 63–75, 2016.
- [12] R. N. Clark, "Spectroscopy of rocks and minerals, and principles of spectroscopy," *Manual Remote Sens.*, vol. 3, pp. 3–58, 1999.
- [13] B. Rivard, N. Harris, J. Feng, and T. Dong, "Inferring TOC and major element geochemical and mineralogical characteristics of shale core from hyperspectral imagery," *AAPG Bull.*, vol. 102, pp. 2101–2121, 2018.
- [14] D. A. Burns and E. W. Ciurczak, *Handbook of Near-Infrared Analysis*. Boca Raton, FL, USA: CRC Press, 2007.
- [15] B. Rivard, J. Feng, D. Russell, V. Bhushan, and M. Lipsett, "Hyperspectral characteristics of oil sand, part 1: Prediction of processability and froth quality from measurements of ore," *Minerals*, vol. 10, no. 12, 2020, Art. no. 1138.
- [16] Y. Mehmani, A. K. Burnham, M. D. V. Berg, and H. A. Tchelepi, "Quantification of organic content in shales via near-infrared imaging: Green river formation," *Fuel*, vol. 208, pp. 337–352, 2017.
- [17] B. Rivard et al., "Bitumen content estimation of athabasca oil sand from broad band infrared reflectance spectra," *Can. J. Chem. Eng.*, vol. 88, no. 5, pp. 830–838, 2010.
- [18] M. Khodadadzadeh, C. Contreras, L. Tusa, and R. Gloaguen, "Subspace clustering algorithms for mineral mapping," in *Proc. Image Signal Process. Remote Sens.*, 2018, vol. 10789, Paper 107891 V.
- [19] V. Rodriguez-Galiano, M. Sanchez-Castillo, M. Chica-Olmo, and M. Chica-Rivas, "Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines," *Ore Geol. Rev.*, vol. 71, pp. 804–818, 2015.
- [20] C. Contreras, M. Khodadadzadeh, L. Tusa, P. Ghamisi, and R. Gloaguen, "A machine learning technique for drill core hyperspectral data analysis," in *Proc. IEEE 9th Workshop Hyperspectral Image Signal Process. Evol. Remote Sens.*, 2018, pp. 1–5.
- [21] L. Tuşa et al., "Drill-core mineral abundance estimation using hyperspectral and high-resolution mineralogical data," *Remote Sens.*, vol. 12, no. 7, 2020, Art. no. 1218.
- [22] C. Riccomini, "O rift continental do sudeste do Brasil," Ph.D. dissertation, Ph.D. dissertation, Inst. Geosci., Universidade de São Paulo, São Paulo, Brazil, 1989.
- [23] S. Bergamaschi, R. Rodrigues, and E. Pereira, "Oil shale from the tremembé formation," *AAPG Search Discov.*, Taubaté Basin, Brazil, 2010, Art. no. 80080.
- [24] J. G. Mendonça Filho, R. B. A. Chagas, T. R. Menezes, J. O. Mendonça, F. S. da Silva, and E. Sabadini-Santos, "Organic facies of the oligocene lacustrine system in the cenozoic Taubaté Basin, Southern Brazil," *Int. J. Coal Geol.*, vol. 84, no. 3/4, pp. 166–178, 2010.
- [25] I. V. Souza et al., "Organic and mineral matter changes due to oil generation, saturation and expulsion processes based on artificial maturation experiments," *Geologica Acta, Int. Earth Sci. J.*, vol. 12, no. 4, pp. 351–362, 2014.
- [26] J. S. Dramsch, "Chapter one—70 years of machine learning in geoscience in review," *Adv. Geophys.*, vol. 61, pp. 1–55, 2020.
- [27] G. M. Smith and E. J. Milton, "The use of the empirical line method to calibrate remotely sensed data to reflectance," *Int. J. Remote Sens.*, vol. 20, no. 13, pp. 2653–2662, 1999.
- [28] M. Beitollahi and S. A. Hosseini, "Using Savitsky-Golay smoothing filter in hyperspectral data compression by curve fitting," in *Proc. Iranian Conf. Elect. Eng.*, 2018, pp. 452–457.
- [29] Z. Shen et al., "Towards out-of-distribution generalization: A survey," 2021, *arXiv:2108.13624*. [Online]. Available: <https://www.sciencedirect.com/science/article/abs/pii/S0010482522003468?via%3Dihub>
- [30] T. T. Guimarães et al., "Kerogen type classification in hydrocarbon source rocks using hyperspectral data and machine learning," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2021, pp. 3633–3636.
- [31] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *J. Mach. Learn. Res.*, vol. 13, no. 2, pp. 281–305, 2012.
- [32] G. R. Hunt, "Spectral signatures of particulate minerals in the visible and near infrared," *Geophysics*, vol. 42, no. 3, pp. 501–513, 1977.
- [33] R. N. Clark, T. V. King, M. Klejwa, G. A. Swayze, and N. Vergo, "High spectral resolution reflectance spectroscopy of minerals," *J. Geophysical Res., Solid Earth*, vol. 95, no. B8, pp. 12653–12680, 1990.
- [34] V. Pokorný, "Ostracodes," in *Introduction to Marine Micropaleontology*. New York, NY, USA: Elsevier, 1998, pp. 109–149.
- [35] F. A. Kruse et al., "The spectral image processing system (SIPS)-Interactive visualization and analysis of imaging spectrometer data," *Remote Sens. Environ.*, vol. 44, no. 2/3, pp. 145–163, 1993.
- [36] F. Zhang, J. Liu, J. Lin, and Z. Wang, "Detection of oil yield from oil shale based on near-infrared spectroscopy combined with wavelet transform and least squares support vector machines," *Infrared Phys. Technol.*, vol. 97, pp. 224–228, 2019.



Tainá T. Guimarães received the M.Sc. degree in environmental engineering sciences from the University of São Paulo—São Carlos Engineering School, São Paulo, Brazil, in 2019. She is currently working toward the Ph.D. degree in applied computing with Unisinos University, São Leopoldo, Brazil.

She is also currently a Researcher with Vizlab | X-Reality and GeoInformatics Lab, São Leopoldo, Brazil, working on geoinformatics, hyperspectral remote sensing, hydrocarbon source rocks, and machine learning.



Lucas S. Kupssinskii received the M.Sc. degree in applied computing in 2019 from Unisinos University, São Leopoldo, Brazil, where he is currently working toward the Ph.D. degree in applied computing.

He is an experienced professional with a demonstrated history of working in both higher and technical education and the software development industry. He is a Researcher with Vizlab | X-Reality and GeoInformatics Lab, São Leopoldo, Brazil, working in applied machine learning. He holds a lecturing position in Computer Science with Pontifical Catholic University of Rio Grande do Sul, Porto Alegre, Brazil.



Milena B. Cardoso received the B.S. degree in geology from the State University of Rio de Janeiro, Rio de Janeiro, Brazil, in 2020 and the M.Sc. degree in applied computing from Unisinos University, Sao Leopoldo, Brazil.

She works as a Researcher with Vizlab – X Reality and Geoinformatic Lab, Unisinos University, São Leopoldo, Brazil, working on geoinformatics, remote sensing, oil systems, classification of rocks, and 3-D reconstruction.



Leonardo Bachi received the B.S. degree in geology from Unisinos University, São Leopoldo, Brazil, where he is currently working toward the M.Sc. degree in applied computing.

He is currently a Geologist Researcher with the X-Reality and Geoinformatics—Vizlab, University of Vale do Rio dos Sinos, Sao Leopoldo, Brazil. He is active in the areas of sedimentary geology, basin geology, petroleum geology, and remote sensing. He has experience in UAV data acquisition and 3-D outcrop modeling.



Alysson S. Aires is currently working toward the M.Sc. degree in applied computing with Unisinos University, Sao Leopoldo, Brazil.

He has been a Cartographer and Surveying Engineer since 2020. He works as a Researcher with Vizlab | X-Reality and GeoInformatics Lab, Unisinos University, Sao Leopoldo, Brazil, as an SfM/MVS Photogrammetry Specialist. His research topics are related to digital outcrop models generation and 3-D modeling of drill core and rock samples.



Emilie C. Koste received the B.S. degree in environmental engineer from Unisinos University, Sao Leopoldo, Brazil, in 2017, and the MBA degree in contaminated site management from the University of São Paulo, São Paulo, Brazil, in 2020.

She is currently working as a Researcher with Vizlab | X-Reality and GeoInformatics Lab, Unisinos University, working especially with hyperspectral remote sensing of the environment.



André L. D. Spigolon received the B.S. degree in geology and the M.Sc. degree in regional geology from the University of Brasília, Brasília, Brazil, in 2000 and 2003, respectively, and the Ph.D. degree in geology from the Federal University of Rio de Janeiro, Rio de Janeiro, Brazil, in 2014, in collaboration with the American Geological Survey (USGS).

He is a Consultant in geology with the Geochemistry Management of the General Management of Research and Development in Innovation, Exploration, and Production, O Centro de Pesquisas e Desenvolvimento da Petrobras (PDIEP/CENPES), where he works as a Consultant in the coordination, management and technical dialogue of internal R&D projects and terms of cooperation with national universities and foreign research institutions in multient projects, in the training of people in geosciences, and in the provision of technical-scientific assistance. He specializes in organic geochemistry and basin analysis, with an emphasis on the source rocks characterization, oils, and gases, as well as kinetic and organic facies modeling.



Luiz Gonzaga Jr. (Member, IEEE) received the M.Sc. and Ph.D. degrees in electrical engineering from the State University of Campinas, Campinas, Brazil, in 1996 and 2005, respectively.

He has held lecturing and professor positions with Unisinos University, São Leopoldo, Brazil, where he cocreated and coheads Vizlab | X-Reality and GeoInformatics Lab. He is currently a member of the IEEE Computer Society, Association for Computing Machinery (ACM), and ACM Special Interest Group on Spatial Information and Special Interest Group on Computer Graphics and Interactive Techniques. His research interests include computer vision, real-time computer graphics, graphics processing unit programming, and immersive visualization with a strong interest toward geoinformatics applications.



Mauricio R. Veronez received the M.Sc. and Ph.D. degrees in transportation engineering from the São Carlos School of Engineering State University, University of São Paulo, Sao Paulo, Brazil, in 1998 and 2004, respectively.

He has held lecturer and professor positions with Unisinos University, São Leopoldo, Brazil, where he cocreated and coheads the Vizlab | X-Reality and GeoInformatics Lab. His research interests include global navigation satellite systems, remote sensing, digital imaging, and immersive visualization with a strong bias on geoinformatics applications.