Feature Extraction from EEG signals for detection of Parkinson's Disease

1st Carolline Angela dos Santos Souza
Health and Biological Sciences Postgraduation Program
Universidade Federal do Vale do São Francisco
Petrolina, Brazil
carolladss@gmail.com

2nd Giovanni Guimarães Viana
Electrical Engineering Undergraduate Department
Universidade Federal do Vale do São Francisco
Juazeiro, Brazil
giovanni.viana30@gmail.com

3rd Bruno Fonseca Oliveira Coelho
Doctoral Program in Electrical and Computer Engineering
Faculty of Engineering of the University of Porto
Porto, Portugal
brunofonsecaoc@gmail.com

4th Ana Beatriz Rodrigues Massaranduba
Health and Biological Sciences Postgraduation Program
Universidade Federal do Vale do São Francisco
Petrolina, Brazil
biamassaranduba@hotmail.com

5th Rodrigo Pereira Ramos
Electrical Engineering Undergraduate Department
Universidade Federal do Vale do São Francisco
Juazeiro, Brazil
rodrigo.ramos@univasf.edu.br

Abstract—The Electroencephalogram (EEG) is a medical tool that captures, in a non-invasive way, electrical signals from the brain activities performed by neurons. EEG signals have been the target of study as a biomarker of Parkinson's disease (PD), where several methods of analysis are applied. The present work aims to evaluate features extracted from EEG signals, through methodologies such as HOS, Haralick descriptors, and Fractal Features, as new biomarkers for PD identification. Data from 50 individuals, available at the Open Neuro repository, who underwent an attentional cognitive task were analyzed. RF and SVM algorithms were employed for the classification of the extracted features. The best accuracy achieved was 79.49% in differentiating between Parkinson's subjects and control subjects using Haralick descriptors and RF classifier, suggesting that these features can identify activations in brain areas caused by dopaminergic medication.

Index Terms—EEG. Parkison's Disease. Features. Haralick. Fractals.

I. Introduction

The most commonly diagnosed degenerative disease is Alzheimer's disease, affecting approximately 24 million people, followed by Parkinson's disease (PD), which affects more than 10 million people [1]. PD is caused by the degradation of dopaminergic neurons in the substantia nigra of the brain, presenting motor and non-motor symptoms. The late diagnosis of PD is a reality for those who have it, since the warning point in relation to PD is given by the evidence of motor symptoms, and these symptoms become more explicit as the disease advances.

Several scientific researches are directed to the discovery of methods of predicting the diagnosis of PD. To this end, an instrument widely used for this task is the electroencephalogram (EEG), a medical tool that captures, in a non-invasive way, electrical signals from the brain activities performed by neurons [2]. Some methodologies based on the concept of human-machine interaction, using Artificial Intelligence (AI), have been developed in recent years in order to help the use of EEG as a discriminating factor in the diagnosis of PD. One of the research lines uses event-related potential (ERP) studies. They generally employ the so-called oddball paradigm, which involves an attention task directed at specific stimuli as opposed to frequent stimuli. With this methodology, it is possible to prove that Parkinson's patients do not habituate to new events over time [3].

To develop a classification system for individuals with PD, the characteristics of the EEG signals are crucial, and consequently, choosing the best tools for extracting these characteristics becomes a fundamental activity. Well-used tools for feature extraction are spectral techniques, such as Higher-Order Spectra (HOS) and textural techniques, such as Haralick's Features and Fractals techniques.

Higher-order spectra, also known as polyspectra, are spectral representations of higher-order statistics, that is, moments and cumulants of third, or higher, order. HOS can capture deviations from stationarity, Gaussianity, or linearity in the signal [4]. In addition to the frequency characteristics obtained from the HOS, significant information can be obtained from the textural characteristics. In this context, Haralick features are widely used, defining the correlation of pixel intensity with its neighbors [5]. Like the HOS, Haralick features have been gaining space in biological information classification,

with examples such as detection of diabetic retinopathy [6], detection of region of interest in mammography images [7], and detection of Covid-19 [8].

Another tool used for textural characterization is Fractal Techniques. These techniques are useful for quantifying self-similarity properties and the main approaches are fractal dimension (FD), lacunarity (LAC) and, more recently, percolation (PERC). Fractal techniques have achieved significant success in quantifying non-Hodgkin's lymphoma [9], classification of breast and colorectal tumors [10] and Alzheimer's disease indicator [11].

To analyze the effectiveness of the feature extraction methods used, classification algorithms were employed to evaluate the performance of the possible biomarkers. In this work, two algorithms were used for this evaluation, namely: Support Vector Machine (SVM) and Random Forest (RF). SVM is a well-known technique for its efficiency in classifying a given dataset into two subgroups. During its training session, it forms a hyperplane or decision boundary plane depending on the training dataset that defines [5] classification criteria. RF is an ensemble learning approach algorithm using multiple decision trees and aims to solve classification and regression problems [12].

II. DATABASE AND METHODS

The goal of this work was to extract frequency and textural features from EEG signals to study and evaluate them as potential biomarkers for the diagnosis of Parkinson's disease.

A. Database

We used the database from the project *EEG: 3-Stim Auditory Oddball and Rest in Parkinson's*, conducted by James Cavanagh, made available in the public repository Open Neuro. This database brings together various EEG data with a focus on brain disease studies, thus democratizing access to this information.

This database compiles data from 50 people, 25 individuals with PD and 25 with no PD belonging to the control group (CTRL) [3]. The group of participants with Parkinson's Disease visited the lab twice, with seven days apart between visits: on one visit the individuals were on medication and on the other visit the individuals were 15 hours off Levodopa, the dopaminergic medication used to treat PD. These conditions are referred to as ON or OFF, respectively. To ensure that there were no significant differences related to academic instruction and neurological pathologies between the ON, OFF, and CTLR groups, the subjects underwent the Mini Mental State Examination (MMSE).

The *Oddball* paradigm was used in the experiment, where three auditory stimuli were applied. The standard stimuli employed were 440 Hz tones (with 70% occurrence), target stimuli were 660 Hz tones (15% occurrence) and distractor stimuli (naturalistic sounds with 15% occurrence) (Bradley and Lang, 1999). All sounds were presented for 200 ms and the task took an average of 12 minutes to complete. In this task, subjects had the goal of signaling every time they heard target

auditory stimuli. At the end of the task, it was identified that all subjects understood the premise of the activity, according to all notifications. During this experiment, the EEG signals were obtained using an equipment with 64 electrodes.

B. Signal Pre-Processing

The pre-processing of the signals was performed using two tools, the EEGLab toolbox and the FASTER algorithm. With the help of the first tool, the eye signal (VEOG) and accelerometer data from the EEG equipment were identified and then excluded. Subsequently, we divided the signal into epochs of ± 2 s length around the auditory stimulus, thus ensuring that the signals present the responses to the stimuli emitted. With the FASTER tool, bad channels, epochs and Independent Components are identified. For each aspect, statistical parameters of the data are calculated, and the metric that defines contaminated data is the ± 3 value, thus obtaining signals with artifacts attenuated by eliminating the effects of spurious elements.

C. Image aquisition

Higher-Order Spectra (HOS) are multidimensional Fourier Transforms of higher-order statistics, i.e., they are spectral representations of high-order moments or cumulants of a signal [13]. There are several motivations for using higher-order spectra for signal processing, the main ones being: the extraction of Gaussianity deviation information, the identification of non-minimum phase systems or reconstruction of non-minimum phase signals, and the detection and characterization of nonlinear properties of signals, as well as identification of nonlinear systems.

The third-order spectrum is commonly referred to as the bispectrum, and from the definition one can understand that the bispectrum is the Fourier Transform of third-order statistics, as shown in Eq. 1.

$$B(f_1, f_2) = E[X(f_1)X(f_2)X^*(f_1 + f_2)] \qquad (1)$$

Where $E(.)$ is the expectation operation, $X(f)$ is the Fourier transform of the data series $x(t)$ and $X^*(f)$ its conjugate.

The bicoherence of a signal is its normalized bispectrum. Thus, the output values presented will always occur between 0 and 1, and is defined by the Eq. 2.

$$bic^2(f_1, f_2) = \frac{|B(f_1, f_2)|^2}{E[|X(f_1)X(f_2)|^2]E[X(f_1 + f_2)|^2]} \qquad (2)$$

Figure 1 shows examples of (a) bispectrum and (b) bicoherence images, which are graphical representations of the output parameters in two dimensions by associating pairs of frequencies using contour lines.

D. Feature Extraction

Haralick [14] proposed a method of image analysis based on co-occurrence matrices that use the spatial dependence of the shades of gray presented in the images. To facilitate the analysis of these matrices, 14 numerical values, called Haralick
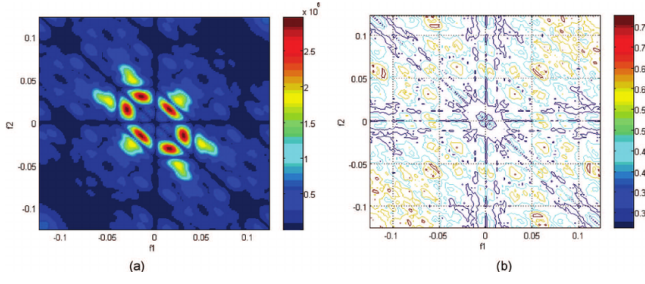
Fig. 1. Bispectrum and Bicoherence plots of heart rate.
CHUA *et al.*, 2009

descriptors, are calculated to help the understanding of the information. Haralick's descriptors are: Second angular momentum, Contrast, Correlation, Variance, Inverse momentum of difference, Sum of mean, Sum of variance, Sum of entropy, Entropy, Variance of difference, Entropy of difference, Correlation measure information I and II, Maximum correlation coefficient.

Note: $p(i,j)$ is the grey level co-occurrence matrix (GLCM).

The second angular momentum is defined as the uniformity between pixels, and is characterized by the Eq. 3:

$$\sum_i \sum_j \{p(i,j)\}^2 \tag{3}$$

Contrast is defined as the measure of the contrast between the intensities of an analyzed pixel and its neighbor, and is characterized by the Eq. 4:

$$\sum_{n=0}^{n_g-1} n^2 \sum_{i=1}^{n_g} \sum_{j=1}^{n_g} p(i,j) \tag{4}$$

For an image with the same gray tone throughout, contrast has a value of 0.

Correlation returns a measure of how correlated a pixel is with its comparison pixel. The range of possible values for correlation is -1 to 1, with a value of 1 for a fully correlated image and -1 for a fully uncorrelated image. The equation that describes this descriptor is given by Eq. 5:

$$\frac{\sum_i \sum_j (ij)p(i,j) - \mu_x\mu_y}{\sigma_x\sigma_y} \tag{5}$$

The variance measures the distortion of the image, using the Eq. 6:

$$\sum_i \sum_j (i-\mu)^2 p(i,j) \tag{6}$$

The inverse moment of difference measures the inverse of the variance, employing the Eq. 7:

$$\sum_i j \frac{p(i,j)}{1+(i-j)^2} \tag{7}$$

Entropy measures the randomness of the elements. It indicates that by containing many null values, the image has little information, and is described by the Eq. 8:

$$-\sum_i p(i,j)log(p(i,j)) \tag{8}$$

The other measures are variations of these.

The gray levels co-occurrence matrix technique consists in having the intensity of each pixel compared to its neighboring pixel in 4 possible directions (0º, 45º, 90º and 135º), as shown in Figure 2. Besides the concept of direction, the concept of distance between pixels is also applied, as seen in the Figure 3.
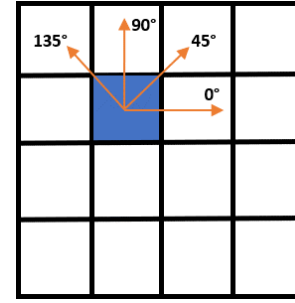

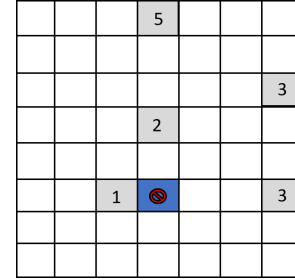Fig. 2. Representation of directions between pixels in an image.

Fig. 3. Representation of distances between pixels in an image.

Image characterization using fractal techniques has been applied to perform the study of shapes that are not defined by Euclidean geometry. In the computational area, some algorithms such as gliding-box [15] are used to compute fractal techniques. These algorithms divide the images into different scales and extract characteristics from each sub-image, thus obtaining local and global characteristics. [16]. A box scrolls through the entire image, to the point of getting information from all pixels. After the task is completed, the size of the box is incremented until it reaches the size determined for completion of the process. The principle of the gliding-box algorithm is represented in Figure 4.
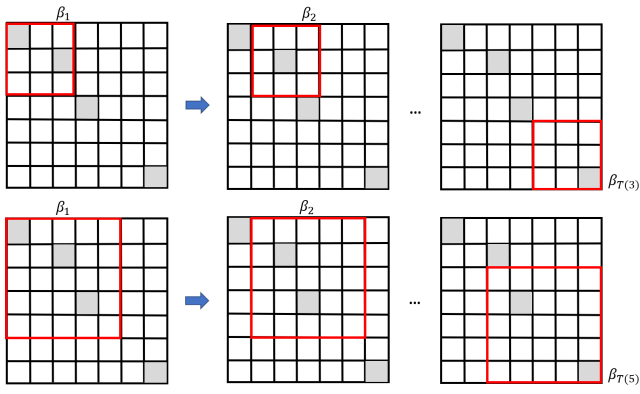
Fig. 4. Visual representation of the algorithm gliding-box.

Each time the box is moved, the fractal dimension (FD), lacunarity (LAC) and percolation (PERC) are analyzed, locally and globally.

Percolation occurs when there is clustering of pixels extending from one end of the image to the other, exemplified by the light green clustering shown in Figure 5:
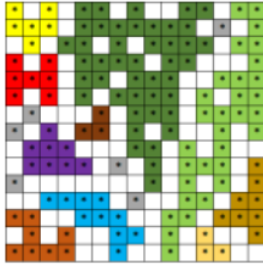

Fig. 5. Pixel grouping.

The fractal dimension is represented by Eq. 9:

$$D(L) = \sum_{m=1}^{L^2} \frac{P(m, L)}{m} \qquad (9)$$

Where: P is the probability that m pixels on a scale L can be labelled as 1.

It is the characterization of irregularity and complexity - spatial extent, propensity to fill space - and is closely linked to the shape of the image. The more "detail" the shape has, the higher the level of fractal dimension it will present. Figure 6 presents shapes that exemplify this concept, with sub-image 1 having a lower level of fractal dimension and sub-image 4 having a higher level.
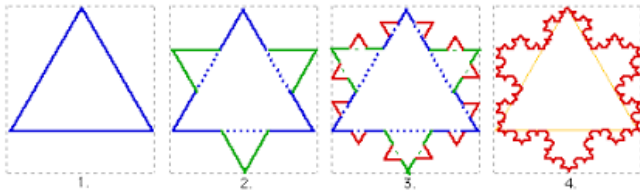

Fig. 6. Forms with ascending fractal dimension level.

Lacunarity is a complementary measure to FD and makes it possible to evaluate the level of spatial filling of the fractal [17], defined by the multidimensional method described by [15], and characterized by Eq. 10 and represented by Figure 7:

$$Lac(L) = \frac{\lambda^2(L) - (\lambda(L))^2}{(\lambda(L))^2} \qquad (10)$$
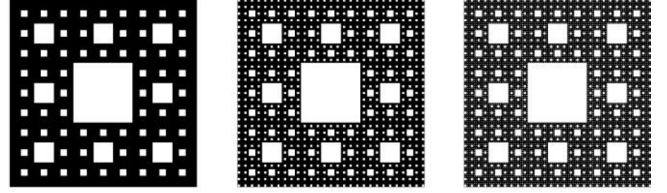
Where: $\lambda(L) = \sum_{m=1}^{L^2} mP(m, L)$

Fig. 7. Three levels of lacunarity.

E. Machine learning algorithms

Machine learning is a subfield of artificial intelligence that deals with computer algorithms that are based on the premise that data analysis guides systems to learn by being able to recognize patterns and assume solutions with minimal human interference. Machine learning algorithms can be classified into three major groups: unsupervised learning, supervised learning, and reinforcement learning.

In supervised machine learning, algorithms are used to induce predictive models by observing a set of labeled objects, usually taken as a training set. With this, a classification algorithm will seek to form a classifier that can generalize the information covered in the training set, with the purpose of later classifying objects whose identification is not known.

An algorithm that stands out within the supervised machine learning method is the Support Vector Machine (SVM), which aims to determine decision boundaries that produce an optimal separation between classes by minimizing errors. Introduced through statistical learning theory by Vapnik (1995), the classification done by SVM is based on the principle of optimal separation between classes, such that if the classes are separable, the solution is chosen so as to separate the classes correctly as much as possible.

Another widely used supervised algorithm for classification purposes is the Random Forest (RF). This model gets its name because it develops a forest of decision trees, which involves repeatedly dividing the data set into two groups taking into account certain criteria, until a predetermined stopping condition is met [18].

III. RESULTS

For each channel, the average of the bicoherences related to its epochs was calculated, that is, the bicoherence of each epoch was obtained, and these results were summed and then divided by the number of epochs. Having this information, it was possible to generate graphical representations for each channel. Since there are 75 subjects (25 ON, 25 OFF and 25

CTRL) and 60 channels, a total of 4500 graphical representations were generated. On these images, two texture feature extraction approaches were applied, Haralick Descriptors and Fractal Techniques.

First approach: Haralick's proposed extraction was used to obtain the texture features of each of the images. Haralick's extraction parameters used were distances between 1 and 10, with unit step, and angles from 0º to 135º, with 45º steps. Using this methodology, 63,000 Haralick features were obtained, being 14 features for each image.

Second approach: In this approach, extraction by Fractal Techniques was used, more precisely with Fractal Dimension and Lacunarization. The box started with side L = 3 and ended with L = 41, this value being incremented by 2. This operation resulted in a matrix with 180 characteristics for each individual, totaling 13,500 fractal characteristics.

Table I shows the average accuracies obtained in Haralick method and Table II shows the results of Fractal techniques.

TABLE I
RESULTS OBTAINED FOR HARALICK FEATURE EXTRACTION.

Algorithm	Group	Angle (º)	Mean Accuracy
RF	ON-OFF	45	66%
	ON-CTRL	135	79,49%
	OFF-CTRL	45	64,15%
SVM	ON-OFF	135	60%
	ON-CTRL	135	75,89%
	OFF-CTRL	45	67,04%

TABLE II
RESULTS OBTAINED FOR FRACTAL FEATURE EXTRACTION.

Algorithm	Group	Mean Accuracy
RF	ON-OFF	42%
	ON-CTRL	73,88%
	OFF-CTRL	66%
SVM	ON-OFF	47,70%
	ON-CTRL	66,30%
	OFF-CTRL	50,25%

To evaluate the performance of the proposed models, the ROC curve was employed. Figure 8 shows the performance of the classifier when classifying individuals from Haralick's features and fractal techniques.
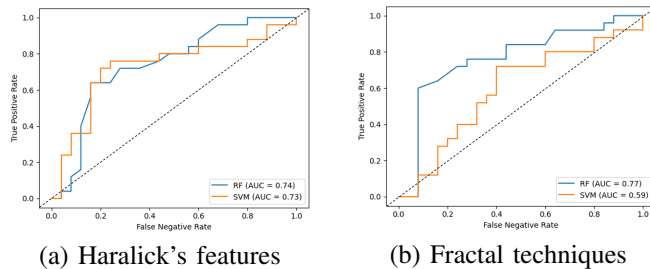

(a) Haralick's features

(b) Fractal techniques

Fig. 8. Roc curves.

IV. DISCUSSION

Observing the two approaches, one can see that the two proposed models work well in discriminating individuals with the disease who are medicated and individuals from the control group, i.e., ON x CTRL. Using only the Haralick features, the best result obtained achieved an accuracy of 79.49% with the Random Forest algorithm. Using the SVM classifier, the best average accuracy obtained was 75.89% for the same group. The results above were validated using the statistical t-test between the mean accuracies against the other pairs of groups (ON-OFF and OFF-CTRL), resulting in p-values ≤ 0.01, which means the significance of the classifications made by the models, i.e., it is certified that the methodology is efficient in differentiating individuals from the ON group and individuals from the CTRL group.

The ROC curves obtained corroborate the accuracy data presented for the models, because they show that the RF classifier curves, in both approaches, have a larger area under the curve than the SVM classifier curves, which is a measure that attests to the accuracy of the classifications instead of their absolute values, as is done with the average accuracy.

V. CONCLUSION

This work investigated EEG signals as biomarkers of Parkinson's disease by employing textural feature extraction. Data from the project EEG: 3-Stimulus Auditory Oddball and Rest in Parkinson's, belonging to the *OpenNeuro* database, were used. Bicoherence, one of the HOS tools, was used to generate images and extract the frequency features, and Haralick descriptors and Fractal techniques were employed to extract the textural features of these images.

The best results were achieved with the Haralick descriptors, which obtained an average accuracy of 79.49% with the random forest classifier, which may be caused by the activation of areas of the brain by the dopaminergic medication that are not normally active in non-parkinsonian individuals.

Analyzing the results obtained, one can see that they diverge from those usually found in papers using EEG signals for the diagnosis of Parkinson's. However, it should be noted that the methods used for feature extraction from EEG signals are unprecedented, which may suggest a new area of exploration.

VI. ACKNOWLEDGMENTS

REFERENCES

[1] M. Erkkinen, K. Mee-Ohk, and M. Geschwind, "Clinical neurology and epidemiology of the major neurodegenerative diseases," *Cold Spring Harbor Perspectives in Biology*, vol. 10, 07 2017.

[2] D. R. Macedo, "Análise espectral de eletroencefalografia para registros patológicos," *Horizonte Científico*, vol. 5, 03 2011.

[3] J.F. Cavanagh, P. Kumar, A. A. Mueller, S. P. Richardson, and A. Mueen, "Diminished eeg habituation to novel events effectively classifies parkinson's patients," *Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology*, vol. 129, pp. 409–418, 12 2018.

[4] K. C. Chua, V. Chandran, U. R. Acharya, and C. M. Lim, "Analysis of epileptic eeg signals using higher order spectra," *Journal of medical engineering & technology*, vol. 33, pp. 42–50, 12 2009.

[5] V. Bhateja, B. Nguyen, N. Nguyen, S. Satapathy, and DN. Le, "Haralick features-based classification of mammograms using svm," *Information Systems Design and Intelligent Applications. Advances in Intelligent Systems and Computing*, vol. 672, 03 2018.

[6] S. Gayathri, A. K. Krishna, V. P. Gopi, and P. Palanisamy, "Automated binary and multiclass classification of diabetic retinopathy using haralick and multiresolution features," *IEEE Access*, vol. 8, 03 2020.

[7] Y. Tasdemir, S.Busra, K. Tasdemir, and A. Zafer, "Roi detection in mammogram images using wavelet-based haralick and hog features," *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 105–109, 2018.

[8] V. Perumal, V. Narayanan, and S. Rajasekar, "Detection of covid-19 using cxr and ct images using transfer learning and haralick features.," *Applied intelligence (Dordrecht, Netherlands)*, vol. 51, pp. 341–358, 2021.

[9] G. F. Roberto, L. A. Neves, M. Z. Nascimento, T. A.A. Tosta, L. C. Longo, A. S. Martins, and P. R. Faria, "Features based on the percolation theory for quantification of non-hodgkin lymphomas," *Computers in Biology and Medicine*, vol. 91, pp. 135–147, 2017.

[10] G. F. Roberto, M. Z. Nascimento, A. S. Martins, T. A.A. Tosta, P. R. Faria, and L. A. Neves, "Classification of breast and colorectal tumors based on percolation of color normalized images," *Computers & Graphics*, vol. 84, pp. 134–143, 2019.

[11] D. Bordescu, M.A. Paun, V.A. Paun, and V.P. Paun, "Fractal analysis of neuroimagistics. lacunarity degree, a precious indicator in the detection of alzheimer's disease," *UPB Scientific Bulletin, Series A: Applied Mathematics and Physics*, vol. 80, pp. 309–320, 01 2018.

[12] M. Sheykhmousa, M. Masoud, H. Ghanbari, F. Mohammadimanesh, P. Ghamisi, and S. Homayouni, "Support vector machine versus random forest for remote sensing image classification: A meta-analysis and systematic review," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 6308–6325, 2020.

[13] C.L. Nikias and J.M. Mendel, "Signal processing with higher-order spectra," *IEEE Signal Processing Magazine*, vol. 10, no. 3, pp. 10–37, 1993.

[14] R. M. Haralick, K. Shanmugam, and I. Dinstein , "Textural features for image classification," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-3, no. 6, pp. 610–621, 1973.

[15] M. Ivanovici and N. Richard, "The lacunarity of colour fractal images," *2009 16th IEEE International Conference on Image Processing (ICIP)*, pp. 453–456, 2009.

[16] G. F. Roberto, L. A. Neves, M. Z. Nascimento, T. A.A. Tosta, L. C. Longo, A. S. Martins, and P. R. Faria, "Ensemble of handcrafted and deep learning features based on fractal geometry for the classification of histology images," *Computers in Biology and Medicine*, 11 2021.

[17] Yuxuan Xia, Jianchao Cai, Edmund Perfect, Wei Wei, Qi Zhang, and Qingbang Meng, "Fractal dimension, lacunarity and succolarity analyses on ct images of reservoir rocks for permeability prediction," *Journal of Hydrology*, vol. 579, pp. 124198, 2019.

[18] M. Schonlau and R. Zou, "The random forest algorithm for statistical learning," *The Stata Journal: Promoting communications on statistics and Stata*, vol. 20, pp. 3–29, 03 2020.