UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

CONNECTIVITIES OF VARIOUS COMPONENTS IN ORGANIC-RICH SHALE

A THESIS

SUBMITTED TO THE GRADUATE FACULTY

In partial fulfillment of the requirements for the

Degree of

MASTER OF SCIENCE

By

Yaokun Wu
Norman, Oklahoma
2019

CONNETIVITIES OF VARIOUS COMPONENTS IN ORGANIC-RICH SHALE

A THESIS APPROVED FOR THE

MEWBOURNE SCHOOL OF PETROLEUM AND GEOLOGICAL ENGINEERING

By

Dr. Siddharth Misra, Chair

Dr. Deepak Devegowda

Dr. Rouzbeh Ghanbarnezhad Moghanloo

Acknowledgements

First, I would like to thank my advisor, Dr. Siddharth Misra for providing me such a good opportunity to enhance my research experience. This thesis can only be made with your guidance and advise during my research study.

Second, I would like to thank Dr. Deepak and Dr. Rouzbeh for your willingness to help me with this thesis review. Your helpful suggestions and recommendations contribute significantly to the completion of the thesis. I also would like to thank Dr. Sondergeld and the IC^3 lab for providing me with the high-quality image data.

I would like to express my sincerest thanks to my family overseas for believing in me. Your love and support for the past two years truly motivate me to pursue my goals.

I would like to thank my officemate Hao Li for giving me suggestions in learning Python programming.

Finally, I would like to thank Na Yuan for taking care of me in the past year.

Table of Contents

Abstract

The physical properties of shale are fundamentally controlled by its microstructure. Connectivity of various components in shale is an important property that governs the transport of mass, energy and momentum. Quantifying connectivity of components is a critical aspect to understand the microstructure of shales. Scanning electron microscope (SEM) imaging technique is a popular technique to capture the microstructure of materials. Before quantifying connectivity of components captured in the SEM image, different components in SEM images need to be identified and segmented. In the first part of this study, an automated SEM-image segmentation workflow involving feature extraction followed by machine learning is developed and tested on SEM images of shale. The proposed segmentation workflow is an alternative to classical threshold-based and object-based segmentation. Four components, namely pore/crack, pyrite, organic/kerogen, and rock matrix including clay, calcite and quartz, are automatically identified and segmented. The performance of the automated SEM-image segmentation workflow, quantified in terms of overall F1 score, on the validation dataset was higher than 0.9. In the second part of this study, five different connectivity-quantification metrics, namely two-point statistical function (S_2), two-point cluster function (C_2), cluster size distribution, travel times computed using fast marching method (FMM), and Euler's number, are tested on SEM images of shale. First, the relationships between the connectivity and the responses of the five connectivity-quantification metrics are determined and validated by statistical analysis on a synthetic dataset of binary images, which contains six types of connectivity from the lowest to the highest. Second, such relationships are directly applied to quantify the connectivity of organic/kerogen and pore/crack components in the SEM images of shale.

Chapter 1. Introduction

1.1 Motivation of the Work

Unconventional reservoirs, especially gas shales have been mostly paid attention to due to the success of hydrocarbon production in the past decade[1]. Shale gas reservoirs have been substantiated to store prolific natural gas[2, 3]. The exploration and production from shales are found to be challenging and expensive as the demand for stable sources increases. Due to their complicated microstructure and extremely low permeability, understanding of the microstructure of shales, petrophysical and mechanical properties of the rocks is a crucial task needed for shale reservoir characterization. The common shale rocks exhibit significant mechanical anisotropy because of the distribution and organization of various minerals[4]. The most direct way of capturing the microstructure of shale is to use image analysis. Limitation in resolution of optical microscopes make observation and analysis of shale rock properties impossible[5]. Using scanning electron microscopy (SEM) technique, stitched mosaic of high-resolution SEM images serves to overcome the limited field of view, which makes high-resolution images perfect for analysis of characteristics in microscale[6]. Fractal geometry, pore structures and heterogeneity characteristics are successfully obtained by data from SEM images[7]. The connectivity of various components in geomaterials governs the transport of mass, energy and momentum. For example, the connectivity of the pore space has critical impact on the shale's unfractured ability to deliver gas to the borehole[8]. However, only limited studies of the connectivity quantification from images are found and no comparative study of connectivity from images is available. It is critical to come up with methods that can quantitatively characterize connectivity and can measure directional and spatial features of connectivity of various components. In this work, we used automated image segmentation techniques for identifying components in SEM images of shale samples, where the components in the study are pores, cracks, organic matter, clay, and pyrite. We tested different metrics for connectivity quantification and applied these metrics to the segmented SEM images in the first step to quantify connectivity of pores/cracks, organic matter in the shale rock sample.

1.2 Organization of the Thesis

The thesis is divided into five chapters and is organized as follows:

Chapter 2 introduces the research background for the study. It includes the background for image segmentation and background for connectivity characterization.

Chapter 3 explains the methodology of machine learning based automatic image segmentation as well as the methodology for connectivity characterization/quantification.

In Chapter 4, SEM segmentation results are shown. The performance of a machine learning model is tested as well as its generalization capability is evaluated. The results from different connectivity metrics are presented and discussed.

In Chapter 5, Conclusion and limitation for this work are presented.

Chapter 2: Research Background

2.1 Image Segmentation Background

Scanning electron microscope (SEM) image analysis facilitates the visualization and quantification of the microstructure, topology, morphology (in the secondary electron mode, not in the backscattered electron mode) and connectivity of distinct components in a porous geological material. The process of the division of an image into spatially continuous, disjoint and homogeneous regions, known as image segmentation, is a crucial step prior to image analysis. Although manual segmentation performed by the subject matter expert is the most reliable approach, it requires considerable time, attention and patience, especially for a large size of the high-resolution SEM images.

Traditional image segmentation is commonly categorized into three approaches: pixel-, edge- and region-based segmentation. Histogram thresholding-based segmentation assigns a certain class label to each pixel depending on a specific range of pixel intensity. Images having single or multiple modal in histograms are generally segmented using this method.[9] However, major limitations of the thresholding method include: (1) it requires accurate determination of threshold values and the ranges of pixel intensity for each component, and (2) it is unreliable when such ranges of pixel intensity for two or more components overlap. Another approach is the region-based segmentation, which is also widely applied on SEM images. This method iteratively splits or merges various regions till all the continuous and homogenous regions are identified in the image. Watershed segmentation is one of most popular region-based method in medical image segmentation [10]. However, challenges in selecting proper seed points during the process make the method prone to over-segmentation or under-segmentation. Moreover, such method is computational expensive and sensitive to noise.

Machine learning (ML) application in petroleum and geoscience has shown rapid progress in the recent years. ML methods are capable of learning mathematical rules derived from large datasets to map features and targets, which make task automation possible[11]. In the upstream oil and gas industry, ML methods have been widely adopted in the subsurface characterization and the subsurface processes forecasting. Rostami et al. [12] used ML models to estimate permeability in heterogeneous carbonate reservoirs. Stacked neural networks were recently used to synthesize dielectric dispersion response of geological formations in the subsurface [13]. CO_2 solubility in oil reservoirs is successfully predicted using ML models based on oil saturation, pressure, oil specific gravity, oil molecular weight, reservoir temperature and bubble point pressure [14]. The in-situ pore size distribution in the subsurface formations is generated using deep and shallow neural network models based on wireline logs, such as gamma ray, resistivity, density, and neutron logs [15, 16].

ML applications for image segmentation tasks are also popularized in the recent years. Two types of machine learning techniques, namely supervised and unsupervised learning, are employed in image segmentation. In supervised learning, a ML model learns a function to map inputs (features) to outputs (targets), where the function is accurately derived and can be later used to predict the desired outputs for new, unseen inputs [17, 18]. Segmentation methods using supervised learning can be divided into two broad categories: pixel-wise classification and object-based classification. Anemone et al. [19] use pixel-wise models with an artificial neural network to recognize spectral features of five different classes of land cover in remotely sensed images for locating potential fossil localities. Bauer and Strauss [20] introduced an object-based method to classify soil cover types into stones, residues, shadow and plants. Deep learning, one of the recently populated ML category, also has applications on image segmentation [21]. One of its typical deeply structured neural networks known as convolutional neural networks (CNN) has its inhered advantages for processing image and thus has been mostly developed in computer vision. CNN learns the filters at various scales to be applied on an image for desired classification or regression tasks. Wu et al. [22] constructed a CNN with an encoder-decoder architecture for semantic segmentation, where the road scene objects, such as cars, trees, and roads, were successfully segmented with reasonable accuracy. Ronneberger et al. [23] applied u-net architecture on biomedical segmentation applications such as neuronal structures detection, cell segmentation, where significant improvement in terms of accuracy is achieved. Due to CNN's capability of capturing localized structures in images, it can achieve the most robust segmentation. However, A major drawback of CNN is that it requires large dataset for training. Preparing a training dataset of a large size and high quality is often a challenge in most of the project. In addition, the training for CNN is time consuming. It often takes days or even month to train a reliable model.

Unsupervised clustering also has been used in image segmentation. Compared to the supervised learning method where training data is required, the unsupervised learning can deal with unlabeled data [24]. Shen et al. [25] introduced an extension to traditional fuzzy c means clustering for the segmentation of T1 weighted magnetic resonance (MR) image of brain tissue to identify white matter, gray matter and cerebrospinal fluid.. Self-organizing map (SOM) is another typical method belonging to unsupervised learning. Ong et al. [26] proposed a two-stage hierarchical neural network for segmentation of color images based on SOM. The unsupervised SOM captures dominant colors of an image to generate color clusters which are fed into second level SOM to complete the segmentation. However, few limitations of SOM include proper selection of the dimension of the map and adjustment and optimization of parameters. Jiang and Zhou [27] combined SOM with ensemble learning to improve the segmentation performance. By setting SOM with different parameters and adopting a scheme for aligning different clusters, a robust segmentation result was obtained. However, a major disadvantage is that manually selection of the numbers of regions is required.

Image analysis has been well adopted in the oil and gas industry. Tripathi et al. [28] estimated permeability from thin-section image analysis based on the Carman-Kozeny model. Budennyy et al. [29] used watershed segmentation and statistical learning on polarized optical microscopic images to study the structure of thin section, where the properties of grain, cement, voids, and cleavage are successfully extracted. Rahimov et al. [30] applied local binary pattern (LBP) for feature extraction to classify 3D sub-sample images into six texture categories and obtain the representative permeability. Asmussen et al. [31] developed a semi-automatic region-growing segmentation workflow for rock images to quantify modal composition, porosity, grain size distribution, and grain contacts. Zhao et al. [32] utilized k-means clustering and principal component analysis (PCA) for the remaining oil classification. Oil film, throat retained oil, heterogeneous multi-pore oil, and clustered oil are successfully differentiated.

In terms of SEM images, various segmentation methods have been proposed by deriving information at nanoscale. Narasimha et al. [33] tested kNN, SVMs and Adaboost models on SEM mammalian cells images to segment mitochondria using text-based features. Good performance of the ML models showed the ML methods can perform close to manual segmentation carried out by an experienced user. Aldo et al. [34] applied CNN on SEM images to segment axon and myelin sheath. The model is well structured and trained with the help of data augmentation. Trained on rat SEM images, the model was able to achieve a pixel-wise accuracy higher than 85%. Hughes A et al. [35] utilize preprocessing, segmentation and object classification techniques for SEM image to streamline nanostructure characterization with the help of Ilastik software. The random walk method combined with the semi-supervised pixel classification precisely classified nanoparticles into singles, dimers, flat and piled aggregate. Tang and Spikes [36] used elemental SEM images of seven different elements from shale samples as input features to segment original images into five components such as calcite, feldspar, quartz, total organic carbon (TOC) and clay/pore. However, the limitations lay in the data acquisition of such elemental SEM images and that clay and pore were not successfully being differentiated.

In this study, we propose a workflow for machine-learning-assisted segmentation of SEM images that will enable an improved characterization of hydrocarbon-bearing formations. The machine learning model can automate the process of segmenting 8-bit grayscale SEM images into four distinct component types, namely, pores/cracks, kerogen/organic, matrix and pyrite components. The proposed model can accurately locate organic/kerogen and pore/crack components in organic rich shales, which is a first of its kind demonstration. Importantly, the efficacy of the segmentation technique in the presence of large noise in the data is tested. Based on feature ranking, the second level of wavelet transform is perceived to be the most important feature apart from Gaussian blur for distinguishing pores and organic matter. We also investigate the precision, recall, and F1-score as metrics to access the performance of the proposed method in inner regions and the transition zones. Furthermore, the effectiveness of our approach is demonstrated in comparison to three other popular segmentation techniques, namely FIJI-assisted segmentation, object-based segmentation, and threshold-based segmentation.

2.2 Connectivity Background

The word connectivity both serve as an intuitive notion and a technique term. There is not a single mathematical definition adopted by the community. However, the connectivity has been defined across multiple discipline. In geomorphology, it is defined as the transfer of sediment from one zone or location to another[37]. In hydrological literature, it refers to the physical connection between different parts of a catchment[38]. In geoscience, the connectivity is related to overall structure of a media and is defined as the proportion of the volume of the biggest geobody to the sum of all geobodies[39]]. No matter how connectivity is defined, all the study demonstrates the importance of the connectivity. It is one of the important properties since it governs the transport of mass, energy and momentum. Quantifying connectivity of components is a critical aspect to understand the microstructure of shales. Standard and widely adopted way does not emerge to measure or to quantify connectivity based on images till now. The percolation theory denotes that process of percolation is the transition from disconnected clusters to a large spanning cluster as the proportion increases. Connectivity is defined in percolation theory as the probability of any two cells belonging to the single percolating cluster, where the probability can be estimated numerically by computing the ratio of the volume of the percolation cluster(the dominate cluster) to the volume of the grid for large, finite grids [40].

However, when the proportion of components cannot reach to the percolation threshold the connectivity is literally null and cannot be estimated accurately.

Euler characteristic, a topological invariant, a number that describes a topological space's shape or structure, has been a scalar indicator of connectivity ,which is calculated as the number of clusters minus the number of holes in the cluster in 2D [41]. However, the major limitation lies in no direction information is involved along which connectivity is measured and it fails when the number of holes is substantially higher than clusters.

Indicator variograms are a measure of spatial continuity at a specific threshold. Multiple indicator variograms capture spatial continuity at multiple thresholds and can thus be used to capture differences in continuity at different thresholds[42]. However, the parameters can only be extracted from indicator variogram based on the natural spatial pattern. No quantitively comparison can be found where those parameters directly related to the connectivity.

The microstructure of two-phase random media has been studied using n-point probability functions back to 1982. The theory proposed that information contained in the microstructure can be captured by a set of n-point probability functions, where the probability of finding a certain subset of n -points in the matrix phase and the remainder in the particle phase is determined [43]. However, performing such n point test is extremely computational expensive, which made it infeasible even on the state-of-the-art computational resources. A lower-order version, known as two point statistical functions (S_2), has been proposed and widely used in characterization of structure and bulk properties of random textures [44]. The S_2 function has been adopted in media reconstruction problem due to its capability of capturing structured information. Such methods can determine the extent to which the original structure can be reconstructed by comparison of similarity between function response in original media and in reconstructed one [45].

Orthogonal directions along which the functions are applied are usually considered [46]. That reconstruction results of using orthogonal direction only are less preferable than that of using four direction suggests the limitation of functions calculated only in two directions, where less structural information is preserved [47].

Reconstruction result obtained by adding diagonal direction in the study suggests the structure information such as connectivity is embedded in the target function along the direction it is calculated and also shows the potential of such statistical function for capturing connectivity information [48].

In this study, the connectivity of component in an image is defined by the responses of different metrics. In the two-point correlation function and two-point cluster function, the connectivity is defined as the probability of having two cluster pixels connected. In terms of fast marching method, the connectivity is defined as the percentage of pixels being reached during the boundary evolution. Euler number serves as a direct indicator of connectivity in this study.

Chapter 3: Methodology

3.1 Workflow of Automated SEM Image Segmentation

3.1.1 Introduction of SEM Map

The high-resolution SEM map is acquired using the FEI Helios Nanolab^{TM} 650 DualBeam™ FIB/SEM machine and FEI SEM MAPS™ software at the Integrated Core Characterization (IC^3) lab. **Fig 3.1** shows the SEM map of dimension of 2058 μm by 260.6μm thin section of a shale rock sample from Wolfcamp formation.

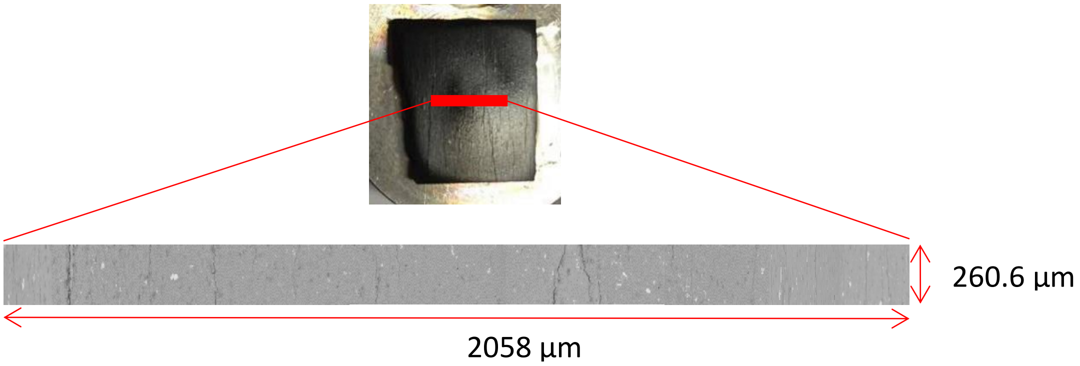


Figure 3.1: High-resolution SEM Map of dimension of 2058μm-by-260.6μm

3.1.2 Workflow

The proposed machine-learning-assisted SEM image segmentation (**Fig 3.2**) is to facilitate the process of identifying the four rock components in the shale reservoir, i.e. whether pixels in a SEM image represent (1) pores/cracks, (2) organic/kerogen (3) matrix comprising clay, calcite and/or quartz, (4) pyrite components. The proposed segmentation workflow involves two steps, feature extraction from images followed by classification of the extracted feature vectors using ML models. To access the performance of models, , the workflow for training and testing stages in chronological order **(Fig 3.2a)** involves (1) pixels selection for training and testing, (2) feature extraction from images, (3) Create training and testing datasets by the compilation of feature vectors of the selected pixels , (4) training ML models using the training dataset, and (5) testing the performance of the ML model on the testing dataset. In the deployment phase (**Fig 3.2b**), the trained model is applied directly on the rest SEM images to obtain the segmented SEM map.

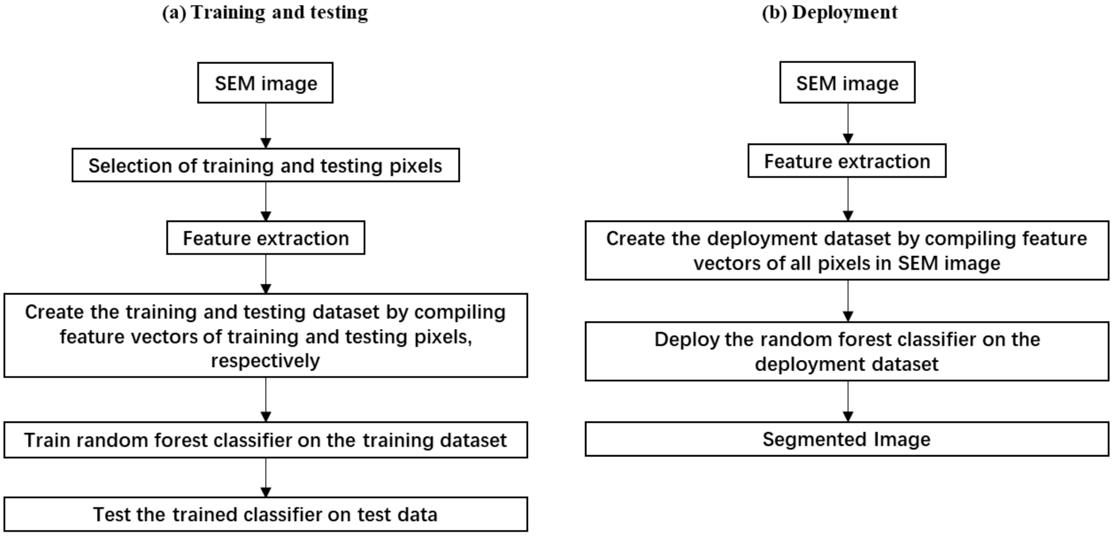

(a) Training and testing
SEM image
Selection of training and testing pixels
Feature extraction
Create the training and testing dataset by compiling feature vectors of training and testing pixels, respectively
Train random forest classifier on the training dataset
Test the trained classifier on test data
(b) Deployment
SEM image
Feature extraction
Create the deployment dataset by compiling feature vectors of all pixels in SEM image
Deploy the random forest classifier on the deployment dataset
Segmented Image

Figure 3.2: Workflows for (a) the training and testing stages for the ML model and (b) the deployment phase for the model

3.1.3 Preprocessing of SEM Map

Because the size of original SEM map is more than a regular computer can handle, the preprocessing is needed in the first place. The SEM map is therefore divided into 1000 same-sized images, where each image has a dimension of 20.58 µm-by-26.06 µm.

3.1.4 Pixels Selection for Training and Testing

Training data is used to fit parameters of ML models. The learning and generalization of a ML model depend largely on training dataset. Good training set selection in the image annotation process can have positive influences on the segmentation model performance while requiring short time to train a model. Pixels selection for creating the training and testing dataset needs to be paid attention to, especially when we deal with pixels around transition area from one component to another. A ML model can be falsified by wrongly annotated pixels.

During the annotation process, ground-truth pixels corresponding to pore/crack were selected from both organic/kerogen region and from the matrix region. In **Fig 3.3,** the rectangles with red-colored edges show where the training pixels were selected. As a result, 705, 2074, 17373, 15000 pixels were selected for pore/crack, organic/kerogen, rock matrix, and pyrite components respectively.

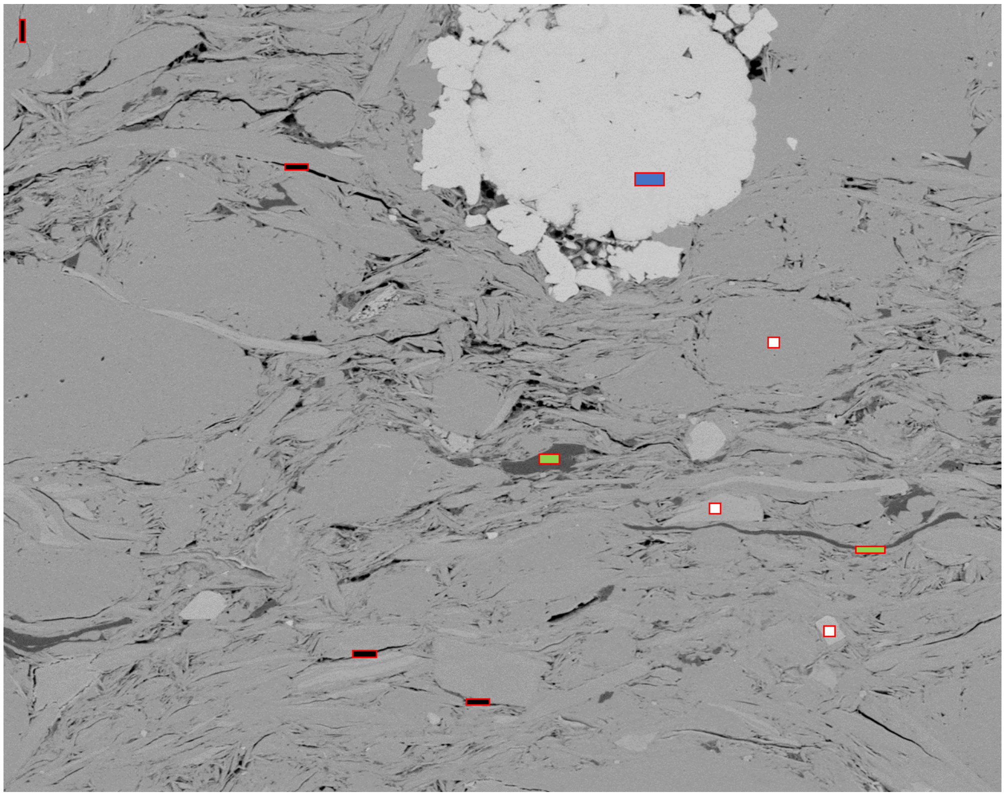


Figure 3.3: Rectangles with red-colored edges indicate the location of training pixels, where green, grey, black, and blue represents the kerogen/organic, matrix, pore/crack, and pyrite components.

It is expected that pores from matrix or inside organic matter can be distinguished. Unfortunately, the segmentation method currently cannot distinguish between pores in matrix and pores in kerogen/organic component.

Test dataset is used to assess the performance of a ML model. Proper selection of testing pixels can reflect the true performance of a model. We divided pixels in the images into two classes based on the location of the pixels, namely, inner region pixels and transition zone pixels shown in **Fig 3.4**.

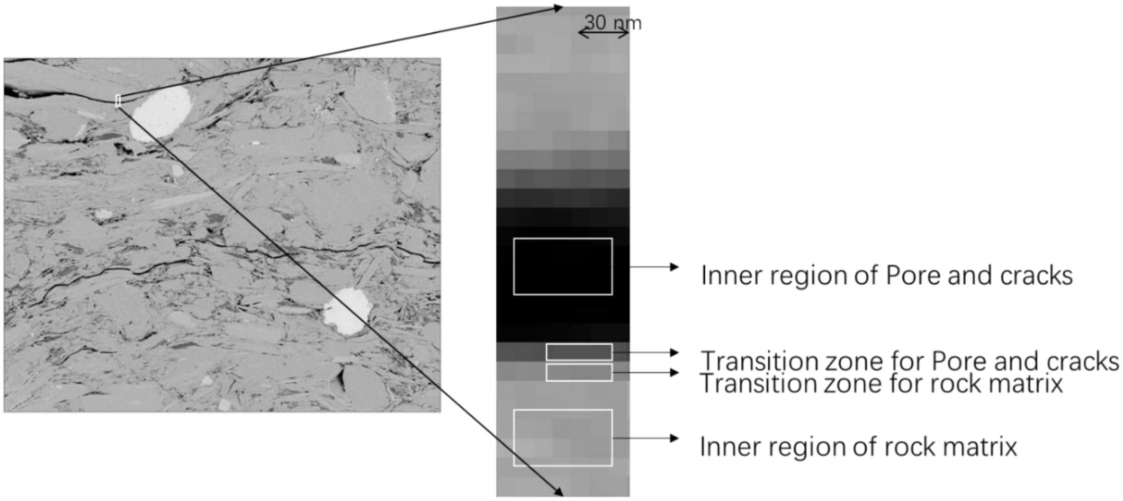

30 nm
Inner region of Pore and cracks
Transition zone for Pore and cracks
Transition zone for rock matrix
Inner region of rock matrix

Figure 3.4: Zoomed in visualization of the inner region (IR) and transition zone (TZ) around crack/pore and matrix interface. Interfaces exhibit grayscale transitions that are hard to segment

A gradual change in pixel intensity can be observed from one component to another. The transition zone for the matrix and pore/crack interface is vague and may seem like organic/kerogen component. It is expected that pixels in transition zone should be more difficult to classify based on the intensity than in inner region. To test the reliability of the model, the test dataset was created with an emphasis on quantifying the performance in the transition zones. Pixels are manually selected from both the inner region and transition zones of the components to constitute the inner-region (IR) and transition-zone (TZ) test dataset, respectively. Manual selection of pixels from the inner region is a straightforward task whereas the selection from the transition zone requires attention to details. The summarization of numbers of pixels **Table 3.1** summarize the number of pixels for each component in inner region and transition zones. The locations where these pixels are selected are shown in **Fig 3.5**.

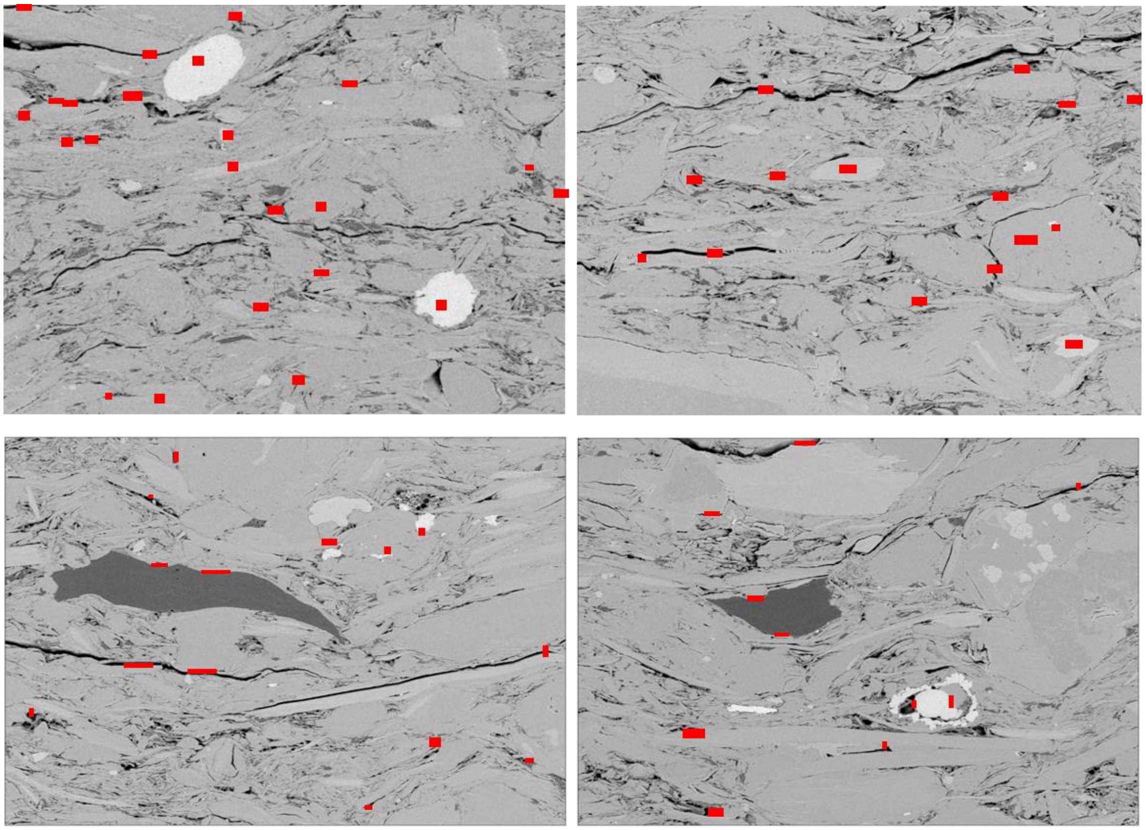


Figure 3.5: Locations of test pixels, which were selected from both inner region and transition zone of different images to effectively test the performance of proposed segmentation.

The red rectangles cover the locations of the test pixels, and the area of each rectangle approximates the number of pixels making up the test data.

Table 3.1: Number of pixels in the test dataset corresponding to the four components in the SEM image

Components	Number of pixels	
	Inner region	Transition zone
Pore & crack	2498	2623
Organic & kerogen	1977	4392
Matrix	2375	3623
Pyrite	1765	3010

3.1.5 Feature Extraction

Intensity of pixels on a gray scale image is a prominent feature to distinguish various components. Obtaining SEM images of uniform intensity for components is a major challenge because the focal distance must be the same throughout the imaging process. Threshold-based method uses only this feature to generate segments. The SEM map, shown in **Fig 3.1**, was used by Tran et al. [6] to identify pores, cracks, organic matter, pyrite, silica-rich clay grains, and calcite-rich clay grains using this method. However, the pixel intensity is sometimes a weak feature when the components to be segmented have overlapping magnitudes of pixel intensity. For our shale images, the threshold-based method has poor performance for distinguishing pore/crack component from organic/kerogen component where pore/crack and organic/kerogen have pixel intensity between 0 to 125, 80-130, respectively., In this case, increasing the number of features is inevitable to ensure robust segmentation result.

Our extensive study indicates that seven categories of features (**Fig 3.6**) are the most important for the proposed segmentation, namely Gaussian blur. Difference of Gaussians (DoG), Sobel operator, Hessian matrix, Wavelet transform, statistical information of the neighboring pixels (local information), and pixel intensity. These features describe each pixel based on the spatial and scale-related information at multiple resolutions. The effectiveness of the features depends largely on the choice of parameters in the corresponding mathematical/statistical transformations. The optimum parameters are selected based on the performance of the ML model on the testing dataset. The descriptions of above-mentioned feature extract technique and the number of features extracted in each category are listed below. Note that the pixel intensity subjects to change when one acquires the image, the study does not consider the variations of pixel intensity range.

Gaussian blur (1 feature)

The feature map of a given image from Gaussian blur is obtained by convolving a 2D Gaussian function with the image. For example, the feature map of the training image from Gaussian blur is shown in **Fig 3.6h**. High spatial frequency information is removed during the process, which result in a smoothed version of original image where noise level in the original image is reduced. A typical 2D Gaussian function is shown in **Equ.1**.

$$G_{2D}(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}$$

where σ is the standard deviation of the Gaussian distribution, x and y are the location indices of pixels in the image. The value of σ determines the extent of the blurring effect. In the proposed method, sigma values ranging from 0.1 to 16 are tested and the sigma value of 3 is determined as the optimum value.

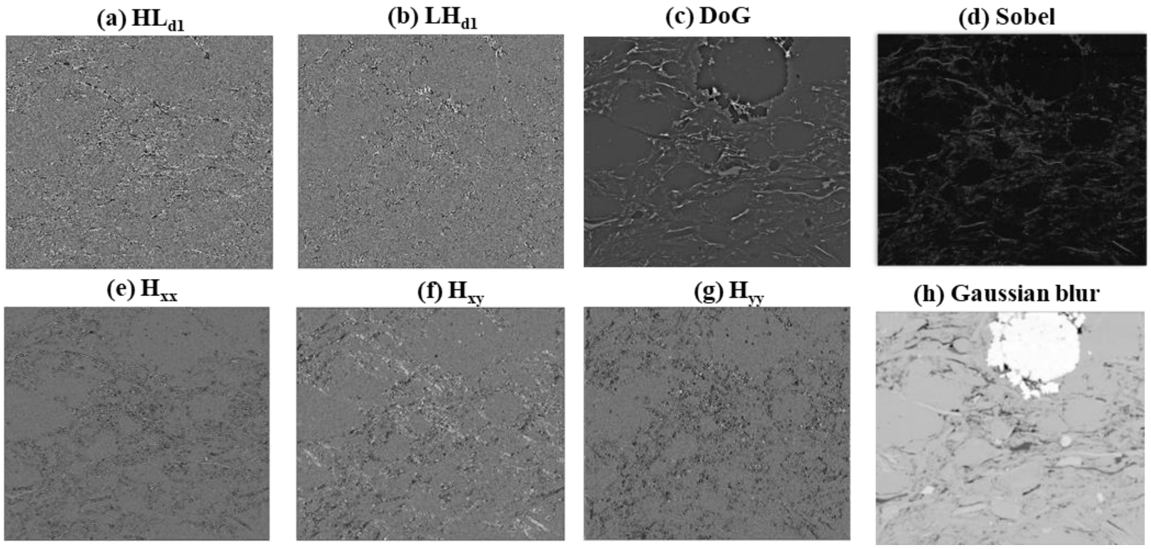


Figure 3.6: **Examples of features extracted from one SEM Image after the first level of processing**

Difference of Gaussians (1 Feature)

Difference of Gaussians (DoG) is calculated as the difference between two feature maps obtained in Gaussian blur with different sigma values. The DoG capture information in a specific spatial frequency domain of original image where such frequency range depends on the sigma values of the two Gaussian blur. Both high-frequency spatial information and low-frequency information are removed during the subtraction of the two Gaussian blur feature maps. This feature extraction technique are popular in object detection, where key points for charactering objects are determined by the response of DoG in an image. **Fig 3.6c** shows the feature map from the DoG where the two sigma values used in the study are 1.414 and 1, respectively.

Sobel Operator (1 Feature)

The Sobel operator performs a 2D spatial-gradient operation on an image for the edge enhancement. The operator consists a pair of 3-by-3 convolution kernels (two perpendicular directions). The two kernels are applied separately to an image to generate the gradients at each pixel. The edges are enhanced due to the sharp pixel intensity changes, where the gradients calculated at pixels around edges are larger than those in the homogeneous region. The feature map of the training image obtained by Sobel operator is shown in **Fig 3.6d**.

Hessian affine region detector (3 Features)

Unlike the Sobel operator for the detection of 1st order variation of pixel intensity, the Hessian affine region detector captures the 2nd order variations of local intensity around a pixel It describes the local curvature of spatial structures in the image; where the shape information is preserved. It has been widely used to structure orientation, brightness detection, and varies structures differentiation. It is computed by convolving an image with the second derivatives of the Gaussian kernel in the *x* and *y* directions. The Hessian matrix H applied on a 2D function $\mathrm{f}(x, y)$ is expressed as

$$\mathrm{H}[\mathrm{f}(x, y)] = \begin{bmatrix} \mathrm{H_{xx}} & \mathrm{H_{xy}} \\ \mathrm{H_{yx}} & \mathrm{H_{yy}} \end{bmatrix}$$

where

$$\mathrm{H_{xx}} = \frac{\partial^2 f}{\partial x^2}\ , \mathrm{H_{xy}} = \frac{\partial^2 f}{\partial x \partial y} = \mathrm{H_{yx}} = \frac{\partial^2 f}{\partial y \partial x}\ , \mathrm{H_{yy}} = \frac{\partial^2 f}{\partial y^2}$$

A standard deviation of 1 in the Gaussian kernel is used in our study. Three feature maps, namely $\mathrm{H_{xx}}$, $\mathrm{H_{xy}}$, and $\mathrm{H_{yy}}$, are obtained and shown in **Fig 3.6e, 3.6f, and 3.6g**.

Wavelet Transforms (6 features)

Wavelet transforms allows multi-resolution space-scale (time-frequency) analysis of signals. Wavelet transform is well known as it can capture both frequency and time/space localization property of the signal being processed. 2D discrete Wavelet transforms generates coefficients with respect to certain basis function (wavelet). In our study, we start off with the Haar wavelet as our basis function and the sensitivity of the ML model to the choice of different wavelet family is compared afterwards. (Haar, filter length of 4 in Dauchies family, filter length of 6 in Coiflet family)

When a single operation (level 1) of the wavelet transform is applied on a given image, four set of coefficients (sub-images) are generated at half the resolution of the original image. Further wavelet transform (level 2 and so on) can be obtained by applying the operation on the one set of coefficients obtained in the previous one. The Level-1 and Level-2 wavelet transforms (decompositions) are shown in **Fig 3.7**. In decomposition level-1, Three sub-images, HL_1, LH_1, and HH_1 are obtained to capture high spatial frequency and local pixel intensity changes in horizontal, vertical and diagonal directions respectively,, whereas LL_1 is an low frequency approximation of the original image The LL_1 can be further decomposed in the next-level decomposition to yield LL_2, LH_2, HL_2, and HH_2 and so on.

In the study, the six high frequency, downscaled coefficients obtained in level 1 and level 2 wavelet transform are inversely used to-reconstruct the horizontal details (HL_{d1} and HL_{d2}), vertical details (LH_{d1} and LH_{d2}), and diagonal details (HH_{d1} and HH_{d2}) of the original image, where the subscripts d1 and d2 represent the level of decomposition. **Fig 3.6a and 3.6b** show the feature maps of horizontal and vertical details. The LL_1, LL_2 and higher-level decompositions are not used in the method due to the following reasons. First, LL_1 and LL_2 are merely approximation (blurred version) of original image, behaving similar to the Gaussian Blur feature and such approximations are not suitable for distinguishing pore/crack from organic/kerogen components, and not for components around interfaces. Importantly as it turns out, the segmentation performance didn't improve with the addition of LL_1 and LL_2. Second, the higher-level decompositions are not preferred because the effect of noise is greatly enhanced.

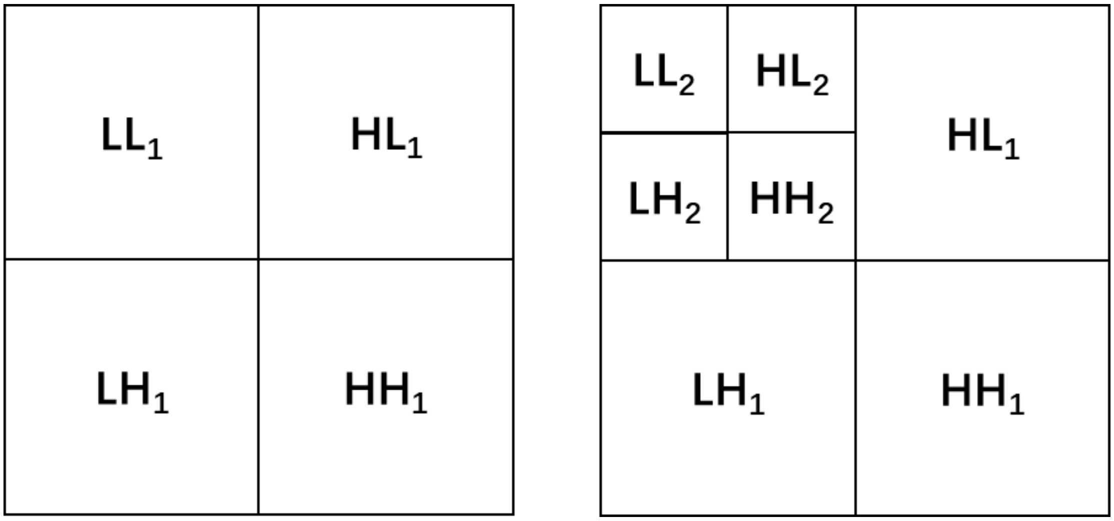


Figure 3.7: Wavelet transforms generated in level one and level two. Each subsequent level generates a downscaled image

Local Information (3 Features)

Local information includes the minimum, maximum and mean values of pixel intensity in a local neighborhood. A 3 by 3 kernel centered at each pixel moves throughout the entire image while the min, max and mean values are calculated at each location of the kernel.

Other Features Investigated for this Study

Features tested but not in use in the study includes empirical mode decomposition (EMD), Local binary pattern (LBP), Scale-invariant feature transform (SIFT) and speeded up robust features (SURF) either due to computational complexity or the lack of reliable computational infrastructure. For example, EMD is a decomposition method similar to the wavelet transforms. Unfortunately, it takes considerable time to run when it was tested on a 256-by-256 image. LBP is popular in texture classification of regions, but not suitable for individual pixels classification. Scale- SIFT and SURF are two other feature extraction methods; However, the two methods specialized only on object detection, and tracking.

3.1.6 Model Selection and Hyper-Parameter Optimization

Tree-based models usually excel in classification problem. The simplest tree-based model, decision tree, always serve as a single unit in ensemble learning due to its overfitting. Tree-based model using ensemble learning includes random forest model, gradient boosting model and Adaboost models. Random forest model is a bagging-type ensemble of decision trees that reduces the variance and bias of the classification task. The representative structure of a Random forest model is shown in **Fig 3.8**. It combines a group of decision tree classifiers trained on various sub-samples of the dataset with bootstrapping. In this study, the random forest classifier is implemented in the Scikit-Learn package, which uses an optimized CART algorithm for building decision trees. The hyperparameters of random forest need to be tuned to overcome the challenge of distinguishing pore/crack component from organic/kerogen component. Important hyper-parameters include maximum depth of the trees, maximum features and the weight assigned to each component. The model selection along with hyper-parameter optimization is achieved through 3-fold cross-validation grid search. Hyperparameters are determined by evaluating the average model performance with different hyperparameters in the cross-validation.

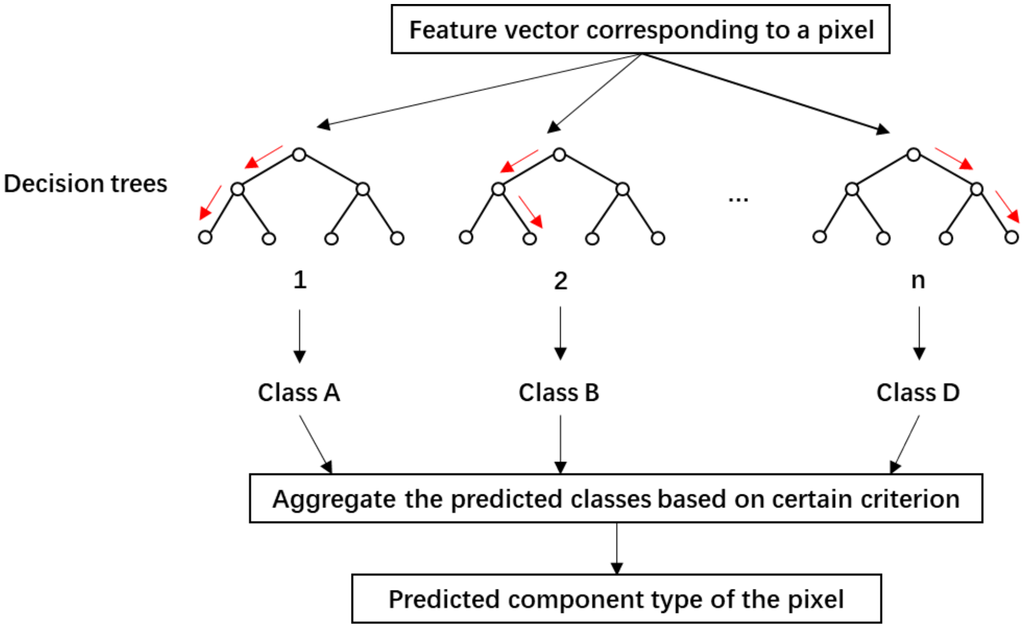

Feature vector corresponding to a pixel
Decision trees
...
1
2
n
Class A
Class B
Class D
Aggregate the predicted classes based on certain criterion
Predicted component type of the pixel

Figure 3.8: A simplified representation of the architecture of the random forest classifier used for the proposed segmentation

The other classification techniques tested in the study include Gradient Boosting (GB), k-Nearest-Neighbor (kNN), Logistic Regression, Linear Support Vector Classifier (SVC), Multi-Layer Perceptron (MLP)). For each unsegmented pixel, kNN first finds *k* pixels, which have feature vectors that are closest to the feature vector of the unsegmented pixel. After that, the unsegmented pixel is assigned a component type that occurs the most among the *k* pixels. kNN requires careful selection of *k*, the number of neighbors. Linear SVC is a binary classifier that finds a boundary that best separates two classes, whereas logistic regression finds a boundary by identifying a log-likelihood distribution that b1est represents the data. Linear SVC and logistic regression require careful selection of parameters: alpha and C that govern the nature of boundary and the penalty of misclassifying few data samples. Non-linear SVC cannot be used for the proposed segmentation because it is inefficient for a large dataset with high-dimensional features. When using neural network model for classification, all features need to be properly scaled and requires hyperparameter optimization with cross-validation to find the optimum values for the regularization term, the number of hidden layers, and the number of neurons in each hidden layer. Based on our extensive study, the random forest model was the most accurate, reliable and computationally inexpensive as compared to others for the desired segmentation. Invariant to the scaling of data and requiring little effort in tuning hyper-parameters while maintaining high reproducibility make the Random forest model the best one in the segmentation task.

3.1.7 Feature Ranking

Feature ranking gives the rank of importance for each feature based on how it contributes to the results. Permutation importance is an operation for determination of feature importance. It replaces one feature at a time with noise data having mean and variance equal to that of the replaced feature. After the replacement, this ranking scheme measures the reduction in the classification score (In this study F1 score is applied). Feature importance is directly proportional to the reduction in score.

3.2 Quantifying Connectivity with Different Metrics

3.2.1 Introduction of the Synthetic Dataset

Performance of the five connectivity-quantification metrics are tested and compared to quantify the connectivity of different components in the SEM images. To that end, the five metrics will be applied on six types of synthetic binary images with different levels of connectivity. The six types of binary images will be referred as Type 1 to 6. Type has the best connectivity of the white component, whereas the Type 6 has the worst connectivity of the white component.

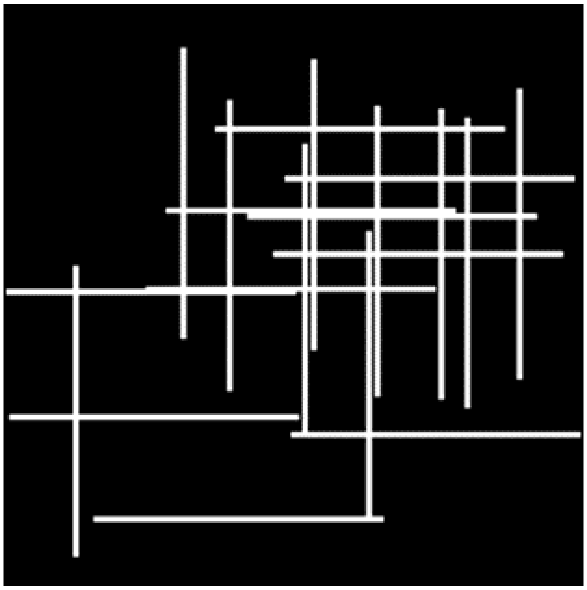


Figure 3.9: A typical binary image of Type 1 connectivity

One typical synthetic binary image of Type 1 connectivity is shown in **Fig 3.9.** The image contains ten horizontal bars and ten vertical bars in white with random distribution. All the bars have the same dimension, i.e. hundred pixels in length and two pixels in width. The dimension of the synthetic binary image are 200 pixels by 200 pixels. White pixels represent the component of interest for which the connectivity is to be quantified, whereas pixels in black represent the background. Type 1 image has approximately 4000 pixels in white representing 10% fraction of the entire image. 500 different realizations of such image are obtained by randomly selecting the location of the bars.

For creating the Type 2 images, all the bars have the same dimension, i.e. fifty pixels in length and two pixels in width. 500 images of Type 2 are generated by random redistribution of the smaller bars. A typical Type 2 image is shown in **Fig 3.10**.

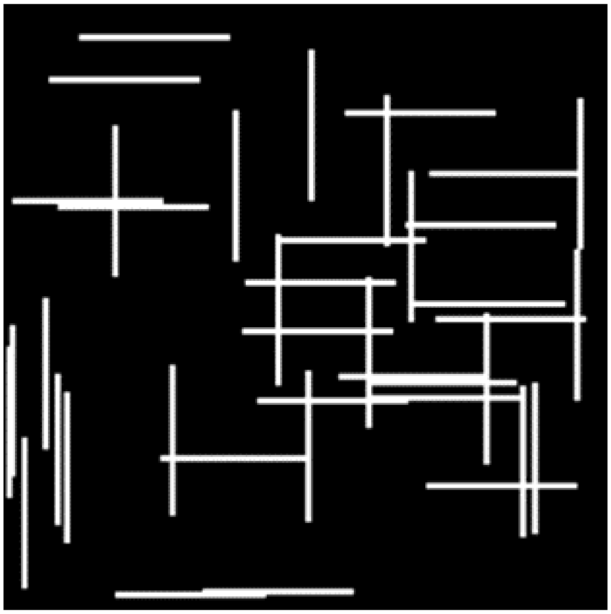
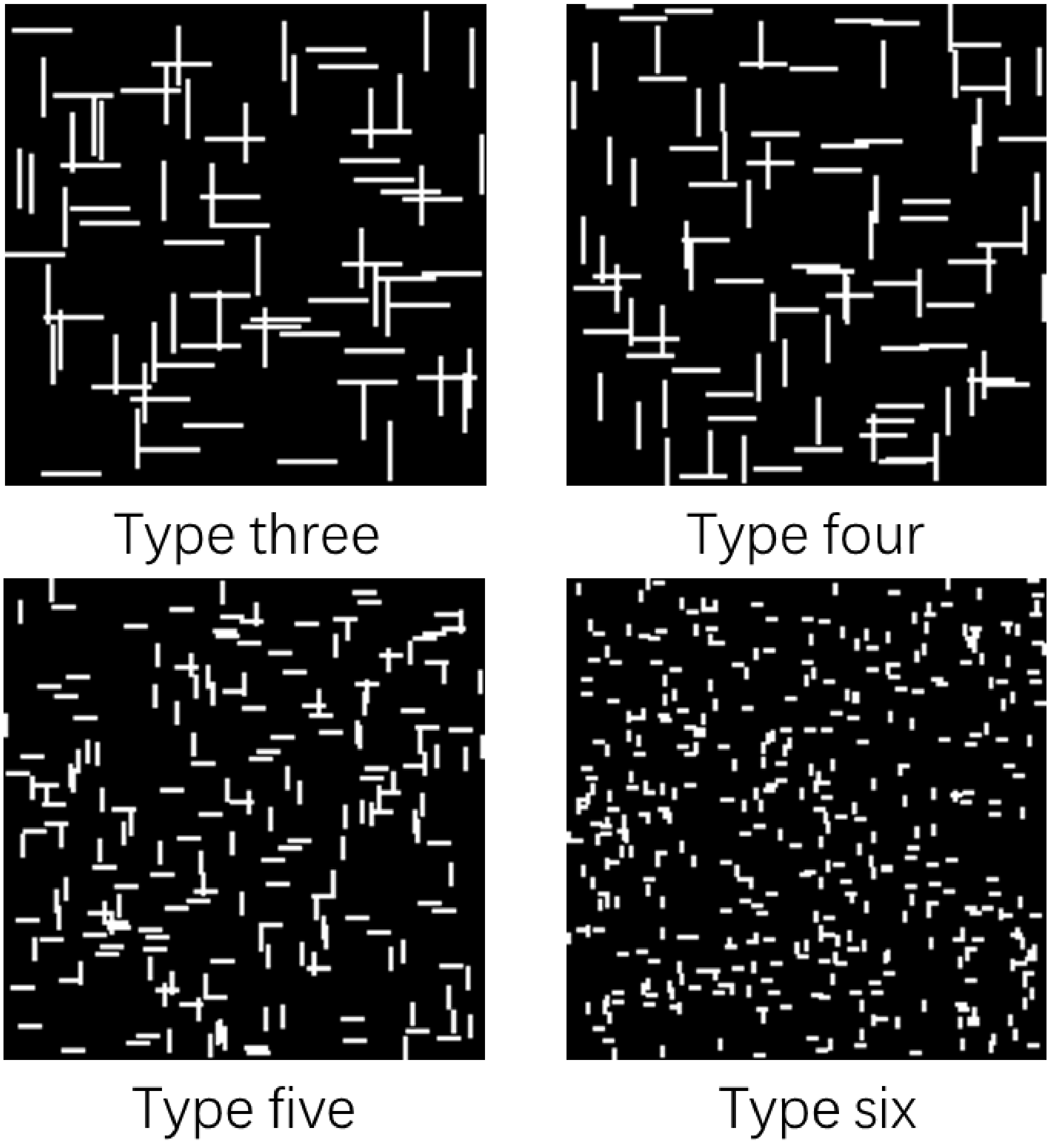


Figure 3.10: A typical binary image of Type 2 connectivity

Smaller length of bars was used to generate synthetic binary images with other four types of connectivity. With the reduction in length of the bar, the connectivity of the white pixels in the binary image reduces. The typical images for these types of connectivity are shown in **Fig 3.11.**



Type three

Type four

Type five

Type six

Figure 3.11: A typical binary image of Type 3, 4, 5 and 6 connectivity

For evaluating the connectivity-quantification metrics, 500 different realizations of randomly distributed bars were generated for each type of connectivity. Each image for each connectivity type has approximately 10% fraction of white pixels. The assumption is that these different realizations for each connectivity type have relatively similar connectivity of white pixels.

3.2.2 S_2 and C_2 Functions

A binary indicator function $I^{(i)}(\boldsymbol{x})$ describes the affiliation between pixels for 2D digitized images [48]. For the synthetic binary images, the indicator function takes the following form at each location $\boldsymbol{x}$ in the two-dimensional Euclidian space:

$$I^{(i)}(\boldsymbol{x}) = \begin{cases} 1, & \boldsymbol{x} \in \boldsymbol{V}_i \\ 0, & \boldsymbol{x} \in \overline{\boldsymbol{V}_i} \end{cases}$$

where $\boldsymbol{V}_i$ is the region occupied by component *i* and $\overline{\boldsymbol{V}}_i$ is the region occupied by the components other than component *i*.

The S_2 statistical function is calculated as the probability of finding two pixels belonging to the same component type separated by a distance of r. There may not be a path connecting the two pixels. The S_2 function consider two pixels belonging to the component type which may be disconnected.

For a certain realization, the probability of two pixels of the same component type at a distance *r* is calculated as the ratio of the number of paired points belonging to the same component type at a distance of *r* to the number of all possible combinations of paired points at a distance of *r*. The paired points are selected randomly for a specified direction. In our study, C_2 and S_2 are calculated along four directions as shown in **Fig 3.12.**

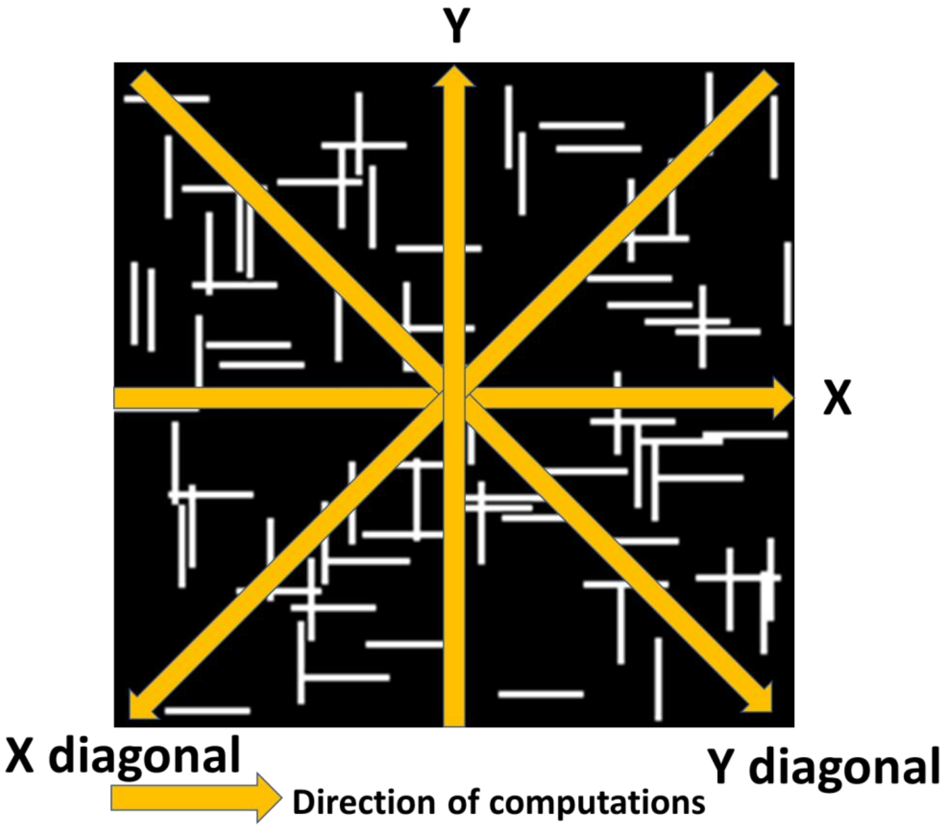


Figure 3.12: Schematic for S_2 and C_2 correlation function computed in four directions (X, Y, X-diagonal and Y-diagonal) for the two-component synthetic binary image

We choose only four directions to calculate the probability because the distance between two pixels in the response of S2 is specified as integer numbers, it is impractical to select paired points at such distance in all directions of 360 degree. It would also be extremely computational expensive if all possible directions are considered.

By definition, C_2 statistical function is different from S_2 in that it requires paired pixels to lie in the same cluster, where a cluster is defined as a group of connected pixels, as shown in **Fig 3.13**. Compared to S_2, C_2 is a better indicator of connectivity since the C_2 consider only two pixels belonging to the same component type where the two pixels are connected.

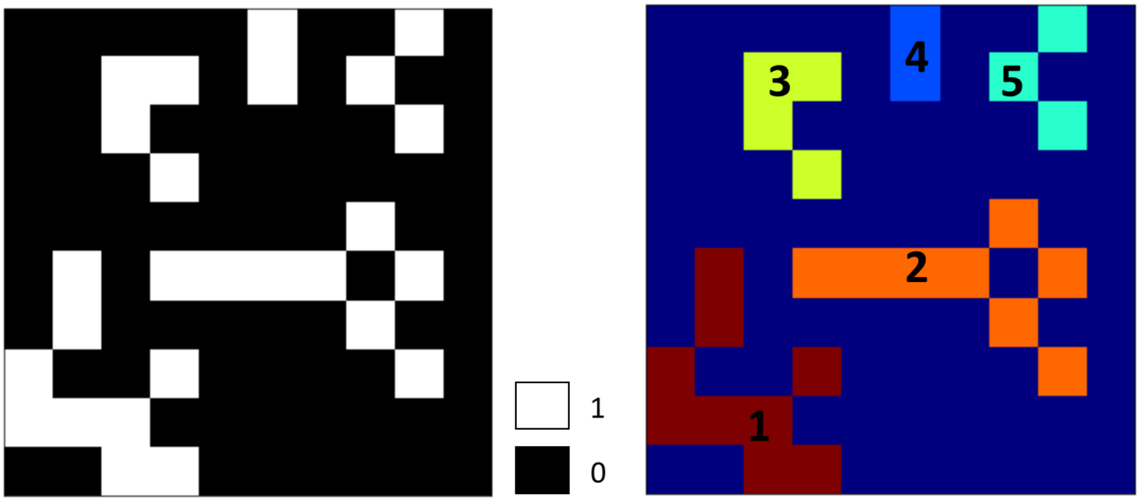


Figure 3.13: Identification of clusters in a sample binary image where five clusters are identified and labeled as 1 to 5.

3.2.3 Fast Marching Method

The fast-marching method is used to model the evolution of boundaries and interfaces. By specifying travel speed for each individual component and the location where the wave start, the travel times from the source point to other pixels (when the contour crosses the pixels) are calculated using the fast-marching computation. For fast marching calculation, the component of interest is assigned a high velocity and the rest of the components are assigned very low velocity. In other words, for the synthetic binary image, prior to fast marching calculations, white pixels were assigned a velocity of 3 m/s and the black pixels were assigned a velocity close to zero. Fast marching computes the travel time for a wave as the wave propagates from the source to other connected white pixels. By randomly initiating travel time calculations from different white pixels in the different realizations for a certain connectivity type, we can obtain a probabilistic distribution of travel times that is related to the connectivity of white pixels. Statistical information contained in the histogram of travel time, as well as the number of pixels being reached and the time of arrival at a certain pixel are considered to be indicators of connectivity.

3.2.4 Cluster Size Distribution

In this study, the number of clusters as well as the size of clusters are considered as indicator of connectivity based on the assumption that connectivity increases with the emergence of large size clusters. Thus, the distribution of clusters size would have connectivity information embedded. A common observation suggests connectivity starts to increase as disconnected points or small clusters merge together given the unchanged quantity of the component before and after. To generate the distribution of cluster size in a 2D image, individual cluster across the image is identified while the size of cluster is calculated as the number of pixels in the cluster.

3.2.5 Euler's Number

Euler's number is a topological invariant. It describes topological space's shape and structure. In 3D, it is the number of clusters minus the number of handles plus the number of holes. It is simplified as the total number of clusters minus total number of holes within clusters in 2D. As the proportion of a component increases starting from zero, at beginning, Euler's number increases due to the increase in the number of clusters and no increase in the number of holes. As the proportion continue to increase and the number of clusters riches to its maximum, the scattered clusters start to merge together, which results in a decrease of the number of clusters and a formation of holes within clusters, which in turn results in a decrease of Euler’s number. Further, holes in the clusters start to be filled up by the component, Euler’s number increases. Eventually, Euler’s number become unity as all the clusters merge together and all holes are filled to form a single large cluster. Thus, Euler’s number serves as a strong, easy-to-understand indicator of connectivity.

Chapter 4: Results and Discussion

4.1 Image Segmentation Results

4.1.1 Four-Component Segmentation

The segmentation model is trained to identify four components: namely, pore/crack (black), kerogen/organic (green), pyrite (blue), and rock matrix comprising clay, quartz, and calcite (light grey). These minerals show differences in grey scale proportional to atomic or bulk densities. The segmentation method involves feature extraction followed by random forest model training to assign a component type to pixels. The proposed method performs better than conventional methods, such as threshold-based segmentation (**Fig 4.1**), object-based segmentation (**Fig 4.2**), and ImageJ Fiji segmentation (**Fig 4.3**).

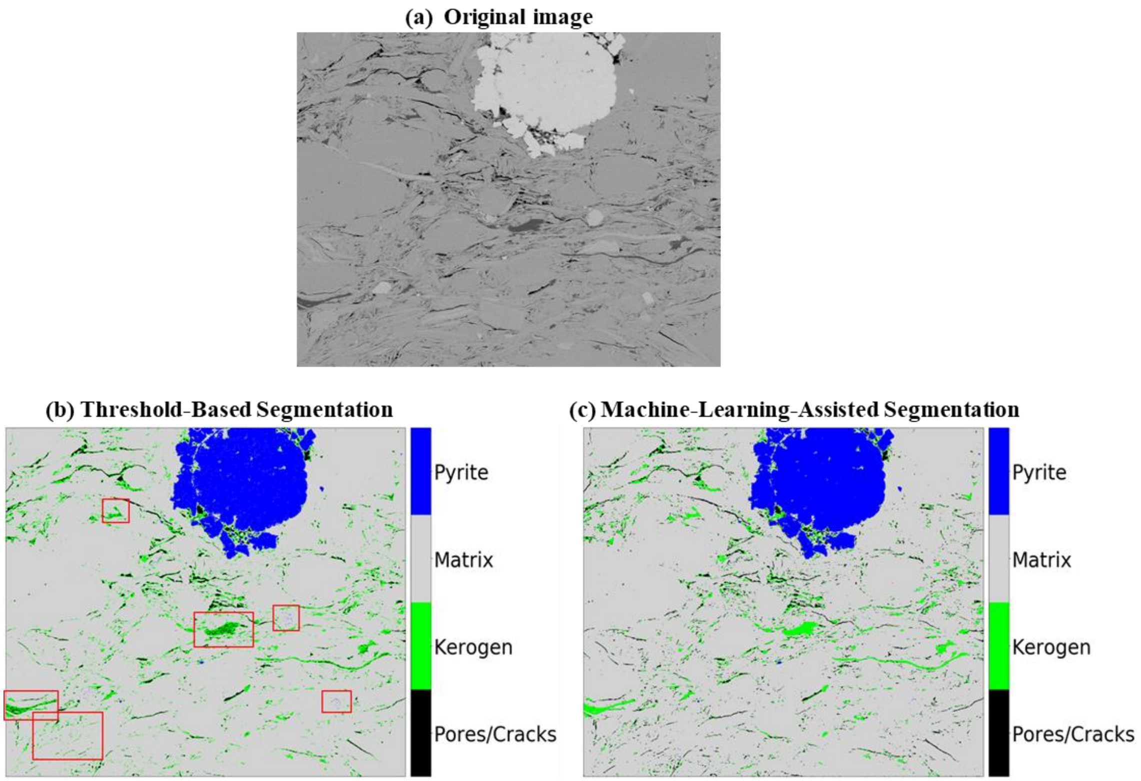
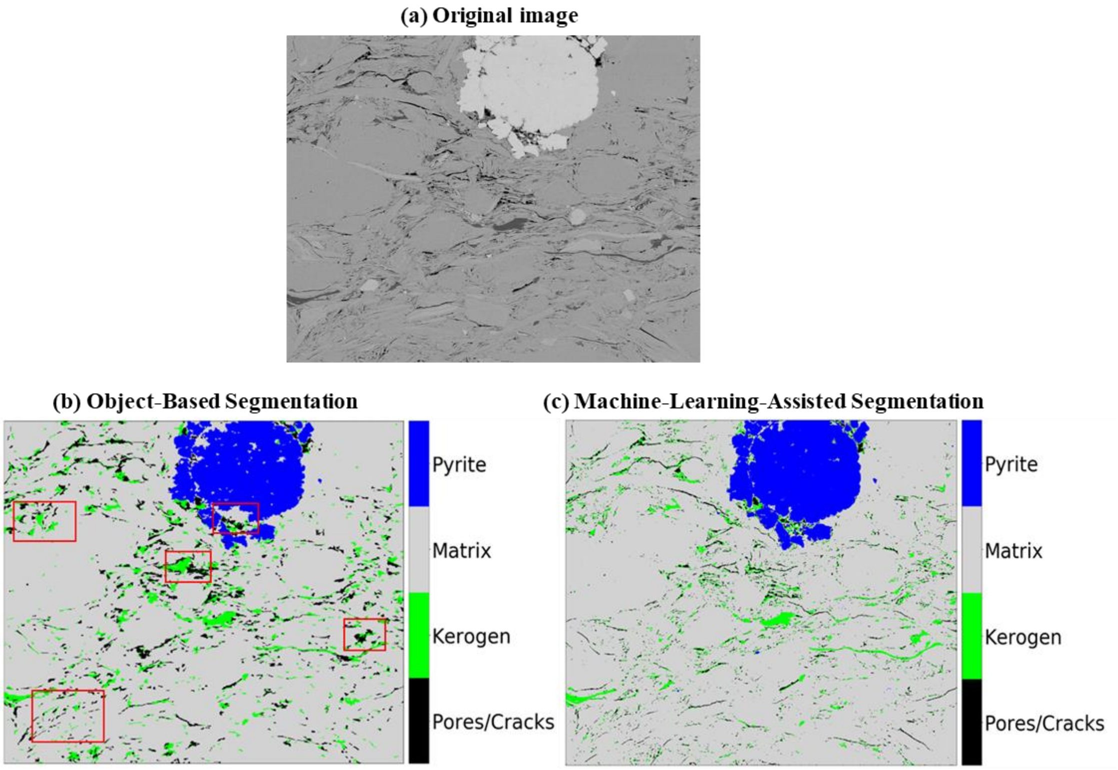

(a) Original image
(b) Threshold-Based Segmentation
(c) Machine-Learning-Assisted Segmentation
Pyrite
Matrix
Kerogen
Pores/Cracks

Figure 4.1: Comparison of SEM-image segmentation generated by (b) threshold-based segmentation and that by the (c) proposed machine-learning-assisted segmentation of (a) original image. Threshold-based segmentation performs poorly in regions indicated by the red-edged boxes.

In the threshold-based method, the pixel intensity range are determined for each component. In the 8bit SEM images, the intensity ranges from 0 to 255. Pixel intensity ranges of 0-80, 81-119, 120-190, 190-255 are manually selected for pore/crack, organic/kerogen, matrix and pyrite components, respectively. A component type was then assigned to each pixel in the image based on the intensity of the pixel. **Fig 4.1** compares the threshold-based segmentation against our proposed method. The threshold-based method performs poorly in the rectangular regions marked with red-colored edges., e.g., the method overpredicts pore/crack by sprinkling pores all over the image and fails to detect it from rock matrix.

Pore/crack and organic/kerogen components tend to be misclassified in threshold-based segmentation due to the overlap of intensity range for the two components.


(a) Original image

(b) Object-Based Segmentation

(c) Machine-Learning-Assisted Segmentation

Figure 4.2: Comparison of SEM-image segmentation generated by (b) object-based segmentation and that by the (c) proposed machine-learning-assisted segmentation of (a) original image. Object-based segmentation performs poorly in regions indicated by the red-edged boxes.

Another popular method widely used in segmentation tasks is the object-based segmentation. It involves object creation, feature extraction, and classification. Unlike pixel-wise based segmentation where individual pixel serve as sample to be assigned label, object-based segmentation firstly create sample (object) as a aggregation of pixels having similar properties, where the aggregation process is defined by a graph-based region comparison algorithm [49]. Then, the statistical parameters of pixel intensity for each sample (object), such as mean, median, minimum, maximum, skewness, and kurtosis, are calculated to describe the sample. Next, these features are combined to form feature vectors and are fed into a ML model for training and testing. A major drawback of the method is that the graph-based region comparison algorithm omits samples having the number of pixels lower than a certain threshold in generating those them, which causes the segmentation result tend to be coarse (**Fig 4.2**). **In Fig 4.2**, pores and cracks spread over a limited number of pixels cannot be identified by the method.

To obtain robust segmentation results, it requires the ML model not be sensitive to the training set selection. Low sensitivity of the model to the training data ensures reproducible segmentation. An image processing package called Fiji is a popular open-source platform for biological-image analysis. One of its plugin called the Waikato Environment for Knowledge Analysis (WEKA) can perform automated image segmentation [50]. The Trainable Weka Segmentation follows the same machine learning workflow for pixel-wised classification. A set of features can be selected from the software such as membrane projection, Gabor filter, entropy and so on. The user defined set of features thus serve as input to varies ML models. The only drawback is that the optimum set of features and ML model are hard to determine, and the segmentation results are sensitive to the training pixels according to our extensive research.

Compared to our segmentation result, the segmentation result from Fiji segmentation varies significantly with different training datasets. As shown in **Fig 4.3**, the Fiji segmentation method frequently misclassifies pore and crack as organic/kerogen matter and pyrite as pores and cracks in the transition zone.

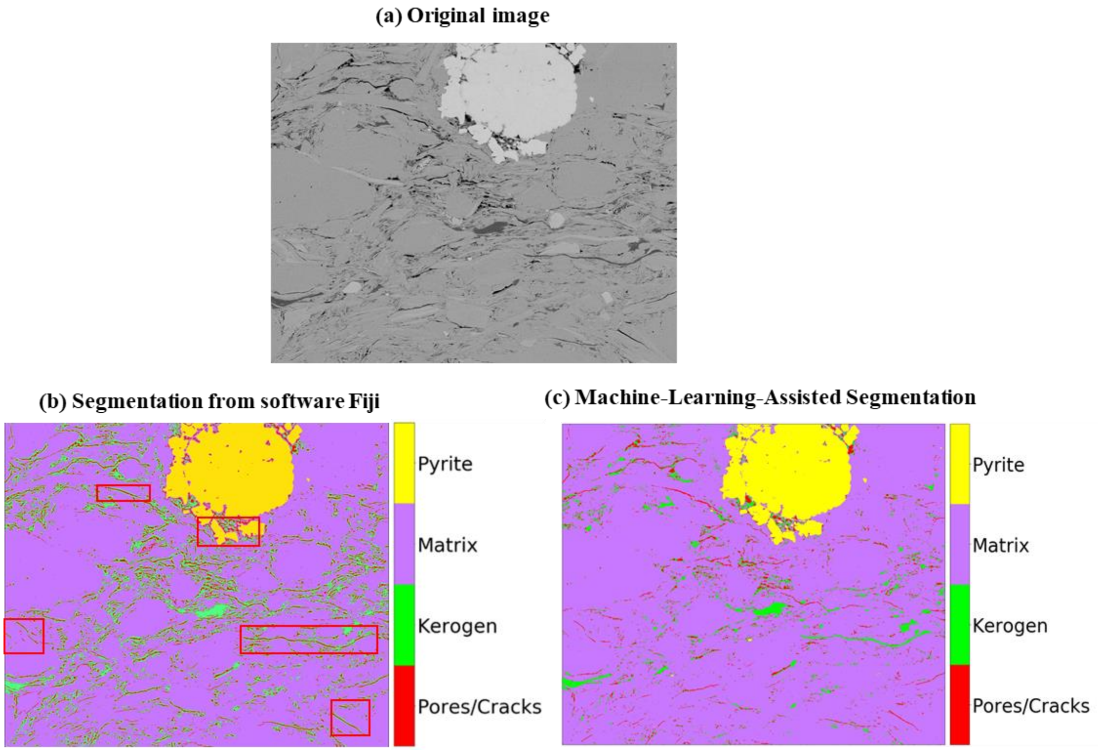


Figure 4.3: Comparison of SEM-image segmentation generated by (b) Fiji-based segmentation and that by the (c) proposed machine-learning-assisted segmentation of (a) original image. FIJI-based segmentation performs poorly in regions indicated by the red-edged boxes.

The four SEM segmentation methods are compared based on their performances on the test image shown in **Fig 4.4.**

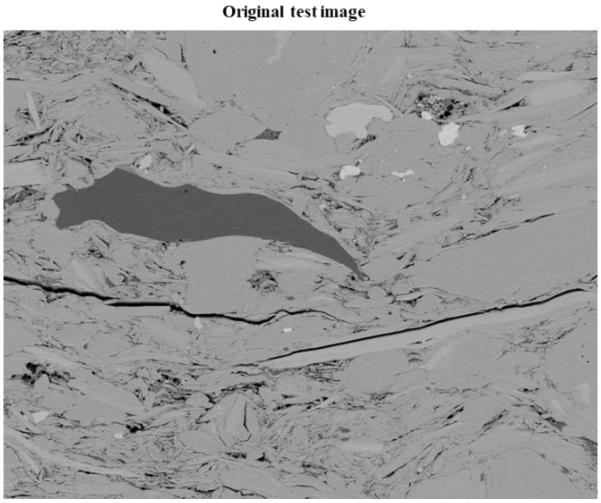


Figure 4.4: One of the **SEM images of shale sample used for testing the four segmentation methods**

The segmentation results comparison is shown in **Fig 4.5**.

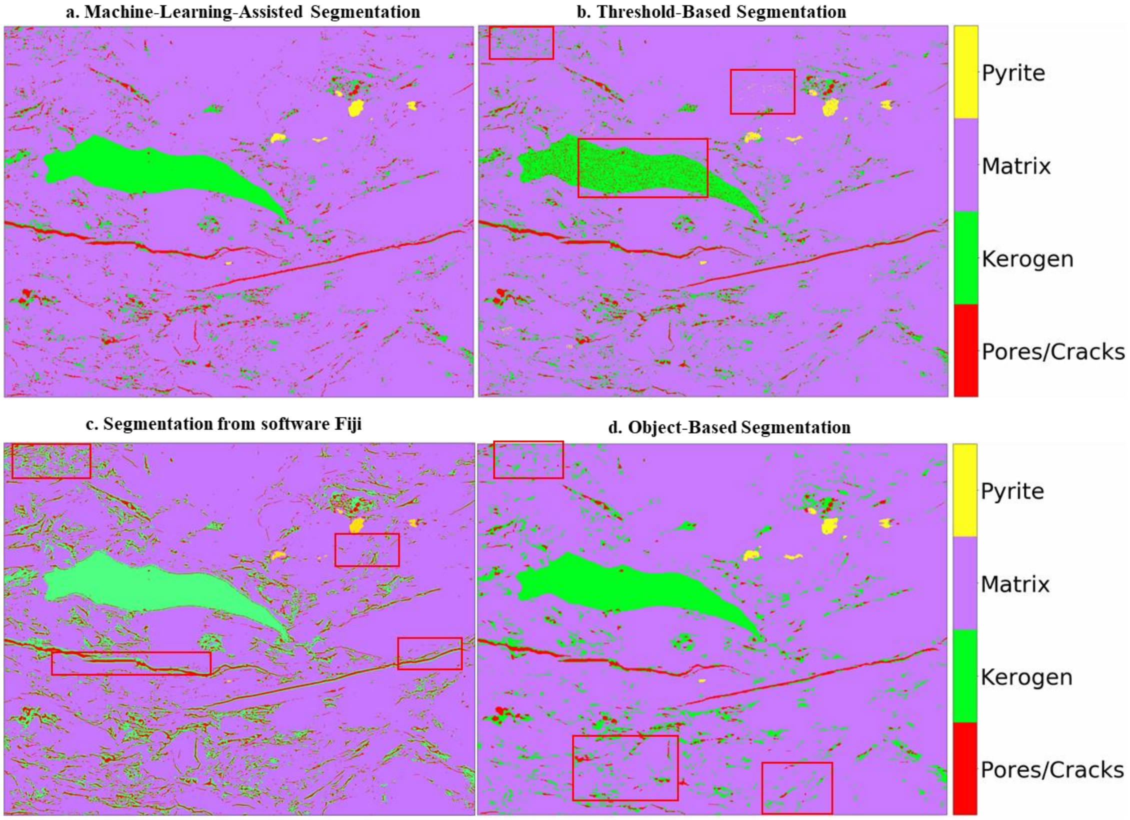

a. Machine-Learning-Assisted Segmentation
b. Threshold-Based Segmentation
Pyrite
Matrix
Kerogen
Pores/Cracks
c. Segmentation from software Fiji
d. Object-Based Segmentation
Pyrite
Matrix
Kerogen
Pores/Cracks

Figure 4.5: Comparison of SEM-image segmentation generated by (a) proposed machine-learning-assisted segmentation, (b) threshold-based segmentation, (c) Fiji-based segmentation, and (d) object-based segmentation. Red edged boxes indicate regions where methods fail.

As observed in the **Fig 4.5b**, the threshold-based method fails for rectangular regions marked with red-colored edges, e.g., pore/crack and organic/kerogen components tend to be misclassified, and the method fails to differentiate pore/crack from rock matrix component. In **Fig 4.5b**, the object-based method fails to identify many pores and cracks spreading over a limited number of pixels cluster. Fiji-based segmentation method misclassifies pixels around interface between pore/crack and organic/kerogen, as shown in

Fig 4.5c. **Fig 4.5d** indicates that our proposed method can identify not only small pores in rock matrix, but also those inside organic matter.

4.1.2 Multi-label Probability-Based Segmentation

Multilabel segmentation is performed using the Random forest model, where the model generates four probabilities of pixels to be one of the four rock components. The probabilities generated by the model describe the confidence in assigning the component types to each pixel. As a result, the uncertainty in the component type assigned by the segmentation is successfully assessed.

Figure 4.6 shows the probability distributions for the four components in a SEM image as obtained by the multilabel model, where the red indicates high confidence and blue indicates low confidence. The segmentation results show that pixels located around the transition zone usually has low confidence associated. Scattered/dispersed pores and organic matter in the matrix also shows region hard to differentiate. The observation confirmed that for each component, regions having high prediction probability usually locate at the inner region of that component whereas uncertainty are observed at boundary region. By selecting a threshold value of 0.7, the probability above which a pixel is assigned to that component type. When none of the single component have a probability greater than 0.7, we assign two labels (component types) to the pixel if the sum of probabilities fortwo components is higher than 0.7.

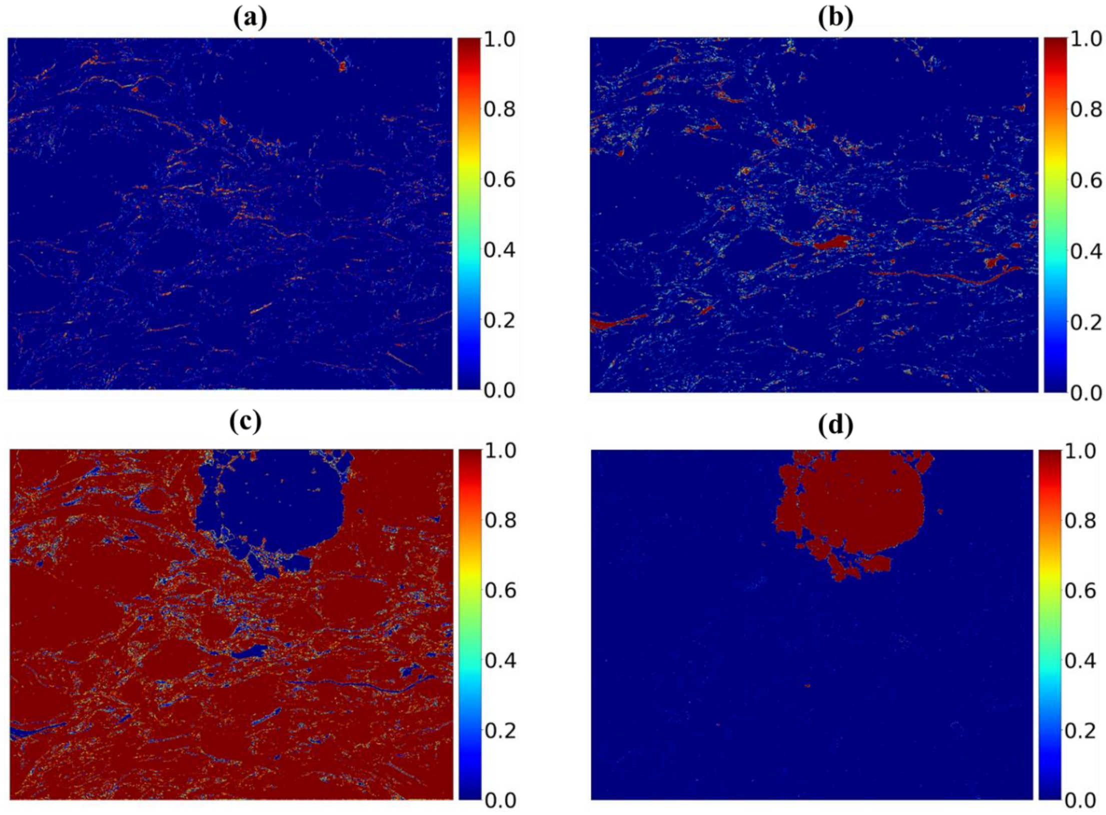


Figure 4.6: Probabilities of a pixel to be (a) pore/crack, (b) organic/kerogen, (c) rock matrix, and (d) pyrite components as generated by the trained random forest classifier for purposes of multilabel classification. Each pixel is assigned four probabilities corresponding to the four components. Regions with probability < 1 indicates the uncertainty in the assigned class-type.

4.1.3 Performance on Testing Dataset

The performance on the test data set is expressed in terms of Precision, Recall and F1 score, AUC-ROC curve and PR curve Precision is the ratio of true positives to the sum of true positives and false positives. Recall (also referred to as sensitivity) is the ratio of true positives to the sum of true positives and false negatives. True positive is when the predicted component of a pixel is the true component of the pixel, whereas false positive is when a pixel is wrongly predicted to be the component of interest. Vice versa, the true negative is when a pixel is correctly predicted to be a component other than the component of interest, whereas false negative is when a pixel is wrongly predicted to be a component other than the component of interest.

The reliability of the component type assigned by the ML model is measured by the precision specific to that component. Similarly, Recall, specific to a component type, is a measure of the classifier's ability to correctly assign that component type; in other words, it is the ability of the model to find the class of interest (similar to the sensitivity of the classifier to a certain class). For example, the scanners at the airport need high recall with respect to dangerous materials but it is not crucial for the scanner to have high precision. The F1 score is the harmonic average of calculated precision and recall. It ranges from 0 to 1, where 0 indicates poor model performance and 1 indicates robust performance. AUC (Area Under the Curve) – ROC (Receiver Operating Characteristics) curve is another way of performance measurement for classification problem at various thresholds settings. ROC is a probability curve and AUC represent degree or measure of separability. It tells how much model is capable of distinguishing between classes. Higher the AUC, better the model is at predicting[51]. The precision-recall curve((PR) is similar to ROC-AUC curve. The PR curve shows the tradeoff between precision and recall for different threshold. A high area under the curve represents both high recall and high precision, where high precision relates to a low false positive rate, and high recall relates to a low false negative rate. High scores for both show that the classifier is returning accurate results (high precision), as well as returning a majority of all positive results (high recall)[52].

As a result, for SEM image with and without twenty percent Gaussian noise, F1 scores of our model are above 0.98 for all the four components in the inner region as listed in **Table 4.1** (without noise) and **Table 4.2** (with noise). Majority pixels in the test images are correctly segmented and the model has good tolerance to noise. The model performance for the transition zone is substantially lower than that for the inner region, especially for the matrix and pyrite components. Matrix component in transition zone has low precision of 0.79 and high recall of 0.9, which indicates that pixels segmented as matrix component have higher uncertainty and the model has ability to identify the actual matrix component correctly. The exact opposite trend is shown for the pyrite component in the transition zone, where a precision of 1 and a recall of 0.74 are observed, which indicates pyrite component is never assigned to any other component whereas pyrite component tends to be wrongly labeled as others.

Table 4.1: Performance of the proposed image segmentation method on the test dataset without noise for the four rock components in the image, where IR and TZ stand for inner-region and transition zone.

Components	Precision		Recall		F1-score	
	IR	TZ	IR	TZ	IR	TZ
Pore & Crack	1.00	0.93	1.00	0.97	1.00	0.95
Organic & Kerogen	1.00	0.96	1.00	0.99	1.00	0.97
Matrix	1.00	0.79	1.00	0.90	1.00	0.84
Pyrite	1.00	1.00	1.00	0.74	1.00	0.85
Weighted Avg.	1.00	0.92	1.00	0.91	1.00	0.91

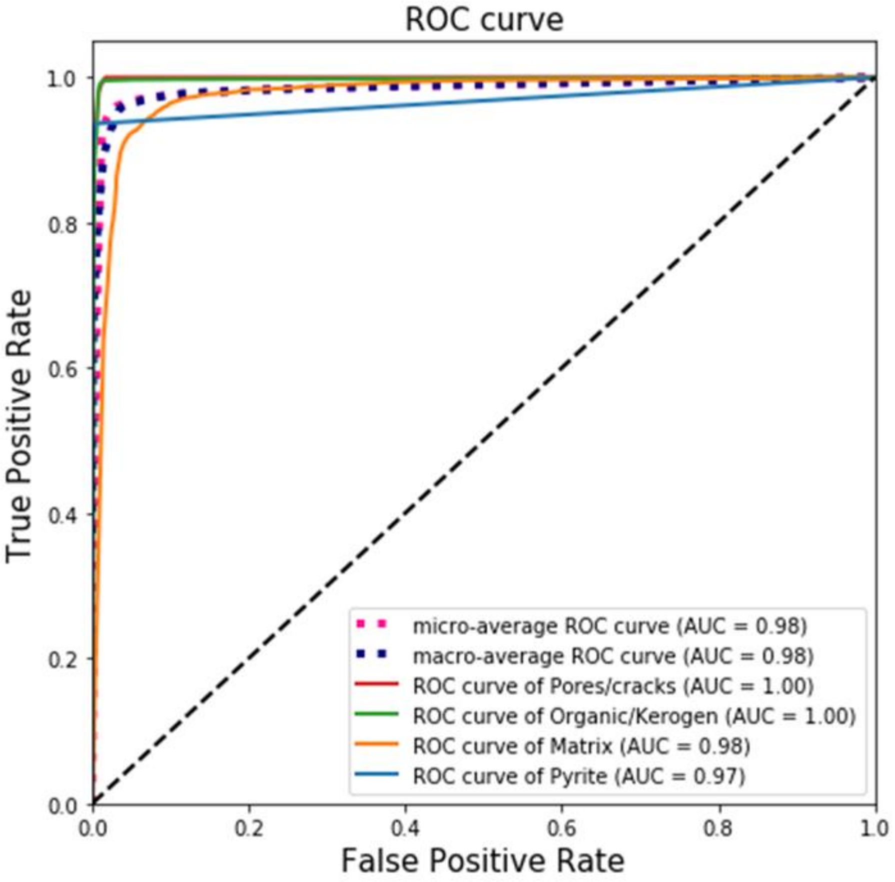


Figure 4.7: ROC-AUC curve for the four components (a) pore/crack, (b) organic/kerogen, (c) rock matrix, and (d) pyrite

Fig 4.7 shows the ROC-AUC curve for the four components. AUC is calculated to be the area covered by the ROC curve with x-axis for the four components, where the area for Pores/cracks, organic, matrix and pyrite are 1.00, 1.00, 0.98, and 0.97, respectively. The high AUC indicates the model perform well for all the four components.

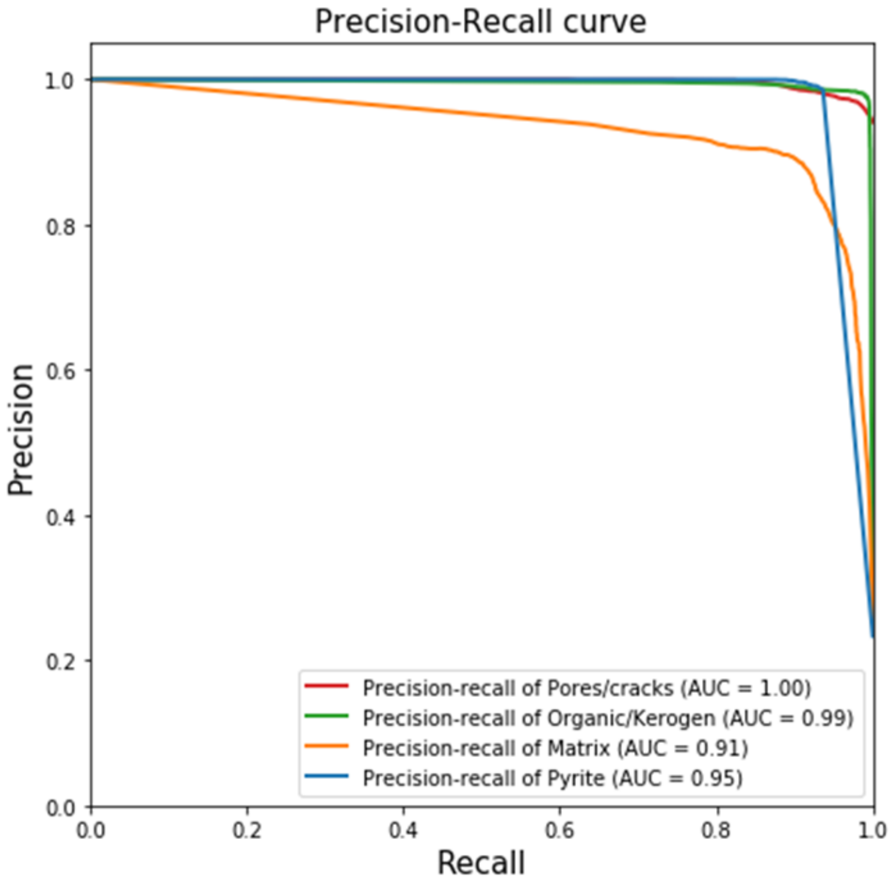


Figure 4.8: PR curve for the four components (a) pore/crack, (b) organic/kerogen, (c) rock matrix, and (d) pyrite

Fig 4.8 shows the PR curve for the four components. In the plot, AUC is calculated to be the area covered by the PR curve with x-axis for the four components, where the area for Pores/cracks, organic, matrix and pyrite are 1.00, 0.99, 0.91, and 0.95, respectively. The high AUC indicates the model performance for pores/cracks and organic matter are better than matrix and pyrite components.

With respect to the transition zone, the F1 scores for pore/crack and organic/kerogen components of noise-bearing test dataset (**Table 4.2**) are similar to those of noise-free test dataset (**Table 4.1**), which indicates that the method is reliable in differentiating pore/crack from organic/kerogen even if the SEM-image has low acquisition quality (i.e. increased Gaussian noise). However, in the presence of noise, the method is not able to segment matrix and pyrite components reliably in the transition zone, where the F1 score drops from 0.84 and 0.85 to 0.75 and 0.79, respectively. The precision for the matrix component and recall for the pyrite component are greatly deteriorated in the transition zone by the addition of noise. The best F1 score is observed for organic/kerogen component in the transition region.

Table 4.2: Performance of the proposed image segmentation method on the test dataset containing 20% Gaussian noise for the four rock components in the image, where IR and TZ stand for inner-region and transition zone.

Components	Precision		Recall		F1-score	
	IR	TZ	IR	TZ	IR	TZ
Pore & Crack	1.00	0.91	0.98	0.96	0.99	0.93
Organic & Kerogen	0.99	0.99	0.98	0.96	0.99	0.97
Matrix	0.97	0.64	1.00	0.89	0.98	0.75
Pyrite	1.00	1.00	0.99	0.65	1.00	0.79
Avg.	0.99	0.89	0.99	0.87	0.99	0.86

Table 4.3: Performance of thresholding-based segmentation method on the test dataset without noise for the four rock components in the image, where IR and TZ stand for inner-region and transition zone.

Components	Precision		Recall		F1-score	
	IR	TZ	IR	TZ	IR	TZ
Pore & Crack	0.85	0.77	1.00	1.00	0.92	0.87
Organic & Kerogen	0.99	0.89	0.78	0.85	0.87	0.87
Matrix	0.98	0.87	1.00	0.82	0.99	0.84
Pyrite	1.00	1.00	0.97	0.86	0.99	0.93
Avg.	0.95	0.88	0.94	0.87	0.94	0.87

Table 4.3 lists the performance of threshold-based method, which is compared with **Table 4.1** to gauge the robustness of the newly proposed segmentation method. Threshold-based method shows good performance only in the inner region of two components, rock matrix and pyrite. For transition zone, A significant drop in performance is observed for pore/crack and organic/kerogen components, whereas an increase is shown for the pyrite component, which is primarily due to the improvement in recall. For both inner region and transition zones, pore/crack exhibits lower precision, whereas organic/kerogen exhibits lower recall.

Table 4.4: Performance of object-based segmentation method on the test dataset without noise for the four rock components in the image, where IR and TZ stand for inner-region and transition zone.

Components	Precision		Recall		F1-score	
	IR	TZ	IR	TZ	IR	TZ
Pore & Crack	0.93	0.59	0.94	0.99	0.93	0.74
Organic & Kerogen	0.97	0.89	0.91	0.71	0.94	0.79
Matrix	0.73	0.50	1.00	0.75	0.84	0.60
Pyrite	1.00	1.00	0.57	0.08	0.72	0.15
Avg.	0.90	0.75	0.87	0.64	0.87	0.59

The performance of object-based segmentation is shown in **Table 4.4.** Low recall and high precision for pyrite component indicates pixels belonging to pyrite component are not reliably segmented. The object-based method performed even worse than the threshold-based method especially for the pyrite and matrix components. Pyrite component has perfect precision for both inner and transition zone. Perfect recall is observed for matrix component in inner region.

Gradient Boosting model trains decision trees in series, where each subsequent tree improves the performance of the previous tree, which leads to reduction in bias with a possibility of overfitting. On the other hand, random forest trains decision trees in parallel with a subset of samples and features, referred as bootstrapping; followed by the aggregation of decisions of the trees. This results in lowering the bias and variance of the classifications. F1 scores for the Gradient Boosting model are similar to those of Random Forest model, as shown in **Table 4.5**. Both precision and recall of the gradient boosting model for matrix and pyrite components are lower as compared to random forest model.

Table 4.5: Performance of Gradient Boosting algorithm on the test dataset without noise for the four rock components in the image, where IR and TZ stand for inner region and transition zone

Components	Precision		Recall		F1-score	
	IR	TZ	IR	TZ	IR	TZ
Pore & Crack	1.00	0.92	0.99	0.98	0.99	0.95
Organic & Kerogen	1.00	0.96	1.00	0.99	1.00	0.98
Matrix	0.98	0.75	1.00	0.89	0.99	0.82
Pyrite	1.00	1.00	0.99	0.68	1.00	0.81
Avg.	0.99	0.91	0.99	0.89	0.99	0.89

4.1.4 Deployment of the Segmentation Model

The trained model is directly applied on other SEM images of the shale sample. For one image of 2058-pixel by 2606-pixel in size, it takes no more than 5 seconds for feature extraction and less than30 seconds is required to obtain the segmentation result. Few random selected segmented images are shown in **Fig 4.9.** The comparison between original and the segmented results clearly outlines the excellent performance of the proposed segmentation methodology.

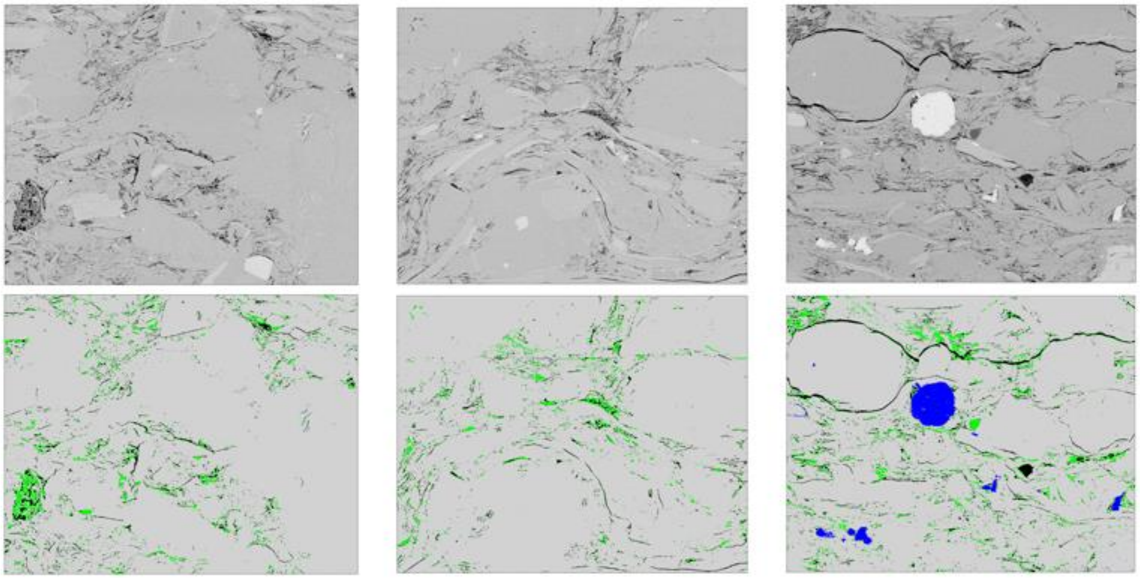


Figure 4.9: Application of the trained segmentation method on other SEM images of shale samples. The segmented images exhibit good consistency when compared to the real images

The porosity (volume fraction of pores and cracks) can be calculated directly from the segmentation results. The porosity is simply calculated as ratio of the number of pixels being pores and cracks to the number of pixels in the image. As the result, the porosity of the image from left to right in Figure 4.7 are calculated to be 2.46%, 1.90%, 3.55%.

4.1.5 Rank of Features

Sixteen features are used in this study. The permutation-importance-based rank of the 16 features from high importance to low importance is reported in **Fig 4.10**.

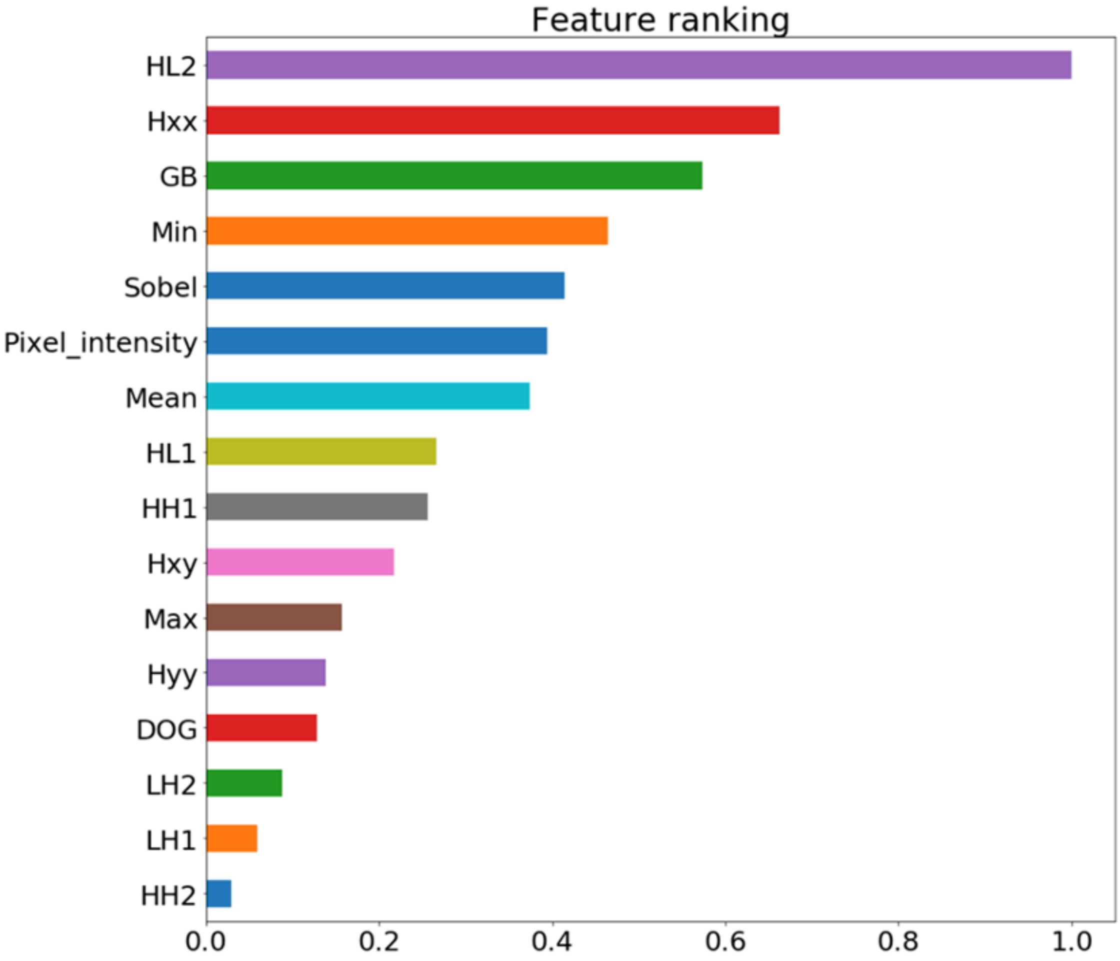


Figure 4.10: Rank of features in the Random forest model based on permutation importance

The rank is as follow from the most importance feature to the least one: HL_{d2}, H_{xx}, Gaussian blur, local minimum, Sobel operator, pixel intensity, local mean, $HL_{d1,}$ HH_{d1}, H_{xy}, local maximum, H_{yy}, DoG, $LH_{d2,}$ $LH_{d1}$and HH_{d2}. The performance of the model constructed by the three top-ranked feature (Gaussian blur, HL_{d2}, and H_{xx}) reduces only 10% of the performance achieved when using all the features, which is a reduction from 0.95 to 0.86 in averaged F1 Score.

4.1.6 Generalization of the Model

This section quantifies generalization of the ML assisted segmentation to a different formation. We study the performance of the model when it applies to testing pixels from a difference formation, SEM map 2. Both the inner-region testing pixels the outer-region testing pixels were selected from different slices of Map-2. We compare the performances of the same model on the inner region testing pixels from the two maps (**Fig 4.11**).

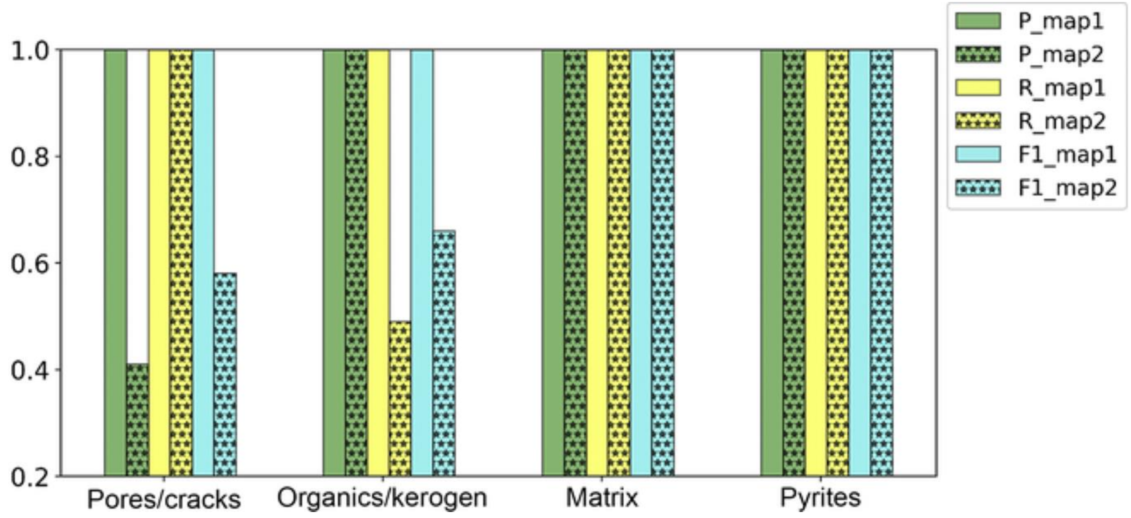


Figure 4.11: Comparison of segmentation model performance (P, precision; R, recall; and F1, F1 score) on inner-region test pixels of Map-1 against those on inner-region test pixels of Map-2. The model was trained on training pixels from Slice 90 of Map-1. Model-1 exhibits good generalization to another formation for the inner regions of matrix and pyrite components.

One thing to note is that there is a significant difference in the distribution of pore/crack components in the two maps and the gray value ranges of each component are different between the two maps. Map-2 is dominated by the presence of pores embedded in organic/kerogen components, whereas Map-1 consists of both organic and inorganic pore systems. In Map-1, the cracks are present in the form of thin strips, whereas Map-2 is characterized by clusters of black pixels representing the pores. As a result, a drop in the F1 score is observed for both the inner and outer-region pixels of the pore/crack and organic/kerogen components, when Model-1 is tested on Map-2 (**Fig 4.11**). For the inner region the precision was 0.41 with a high recall for the pore/crack component, and the recall was 0.49 with high precision for the organic/kerogen component. As supported by the confusion matrix (**Fig 4.12**), a large number of pixels (1615 pixels) belonging to the organic/kerogen in Map-2 are being classified as pore/crack by Model-1, thereby resulting in low precision for pore/crack and low recall for organic/kerogen. Matrix and pyrite components are robustly segmented both in terms of precision and recall. One explanation is that the difference in pixel intensities of pore/crack and organic/kerogen is much smaller than that between these components and the matrix or pyrite components.

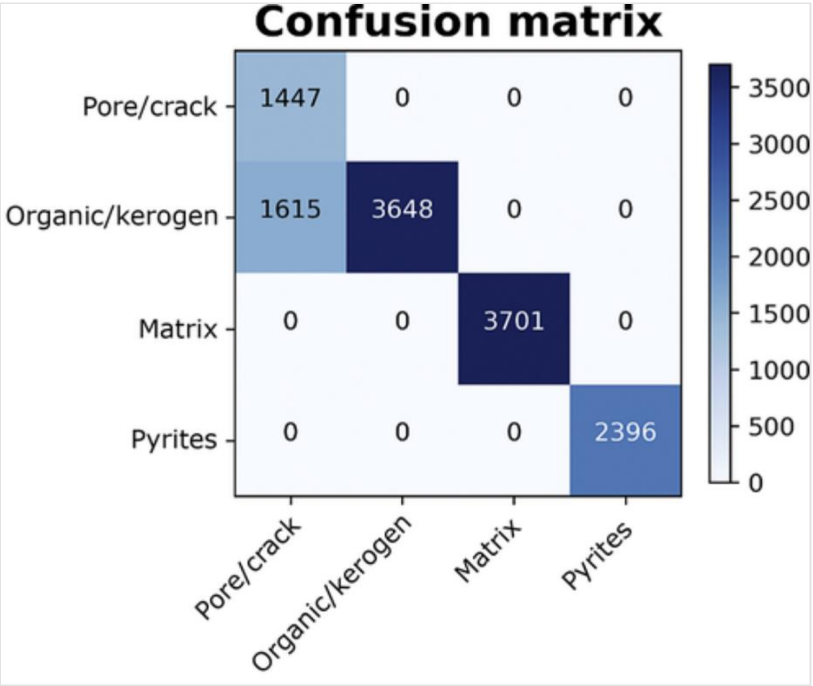


Figure 4.12: Confusion matrix related to the segmentation performance of the model trained on Slice 90 of Map-1 when applied on the inner-region pixels of Map-2. 1615 out of 5263 organic/kerogen pixels got segmented as pore/crack pixel, resulting in a drop in precision of pyrite component and a drop in recall of organic/kerogen pixel.

In a confusion matrix, the diagonal elements represent the number of cases where the true label is same as the predicted label (i.e., true positives), whereas the off-diagonal elements show the number of cases where the components have been misclassified by the model (true negatives and false positives). Therefore, the higher the diagonal values, the better the accuracy of the model. In Fig 4.12, for the matrix and pyrite components, the number of support pixels are equal to the number of diagonal elements, thereby proving that they have been correctly classified. But a significant number of support pixels in organic/kerogen phase has been classified as cracks, resulting in a low value of the F1 score for these two components.

For the outer region, the model was tested on 395, 722, 693, and 2015 pixels corresponding to the pore/crack, organic/kerogen, matrix, and pyrite components, respectively, of Map-2. On an average, the model delivered a lower performance for the outer-region pixels, with F1 scores of 0.89 and 0.81 for Map-1 and Map-2, as compared with that of the inner-region pixels, with F1 scores of 1.00 and 0.82 for Map-1 and Map-2 (Fig 4.13). This occurs since the model tends to misclassify the organic/kerogen pixels as pore/crack because the gray-scale intensities of the components have greater overlap in Map-2. For Map-1 (**Fig 4.13**), we observe much lower precision for matrix and much lower recall for pyrite compared with others, suggesting that the pyrite pixels at the boundary of matrix and pyrite may have been classified as matrix. However, in Map-2, organic/kerogen exhibits very low recall indicating Model-1 is not suitable for organic/kerogen detection. At the same time, the precision for pore/crack of Map-2 is very low, indicating a possibility that the organic/kerogen pixels at the interface of organic/kerogen and pore/crack are being segmented as pore/crack. Interestingly, segmentation performance for matrix and pyrite components improve for Map-2 as compared with Map-1, primarily, due to the shaper contrast at the interfaces in Map-2.

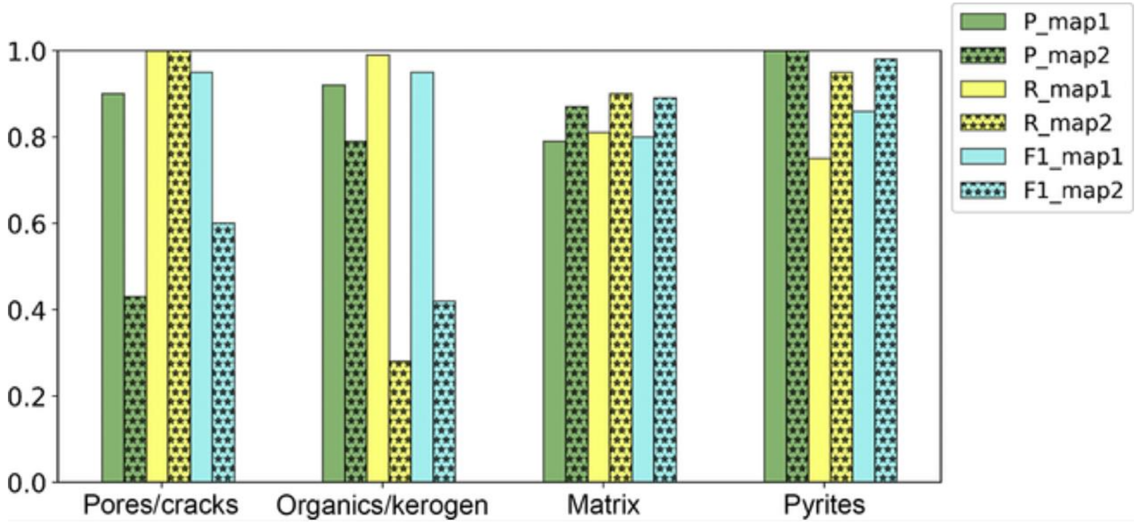


Figure 4.13: Comparison of segmentation model performance (P, precision; R, recall; and F1, F1 score) on outer-region test pixels of Map-1 against those on outer-region test pixels of Map-2. Model-1 exhibits good generalization to another formation for the outer regions of matrix and pyrite components.

4.2 Connectivity Results

4.2.1 S_2 and C_2 Function Results for Synthetic Dataset

The goal is the test the five connectivity-quantification metrics on synthetic binary images of 6 connectivity types. We constructed 500 random realizations for each connectivity type. Following that, the five metrics were applied on the 3000 synthetic images. The calculation of S_2 function are conducted in four directions, two orthogonal and two diagonal direction. At each direction, the probability of two pixels located at a distance *r* to belong to the same component is calculated at the distance *r* ranging from 0 to the maximum length of the image. The size of the synthetic binary image is 200 pixels by 200 pixels; therefore, the largest *r* is set to be 200. Each random realization has its unique probability responses across the four direction. We used the averaged probability across the 500 realizations of each connectivity type at each distance *r* to obtain the representative response. Moreover, the range of two standard deviation from the averaged value at each distance is used to capture the variability of probability for the 500 realizations. Since the bars used in generating these random realizations are either horizontal or vertical positioned, the connectivity in x direction and y direction consider to be the same. Also, the connectivity in *x* diagonal and *y* diagonal considered to be the same as well. S_2 probability for the six types of synthetic images is shown in **Fig 4.14**.

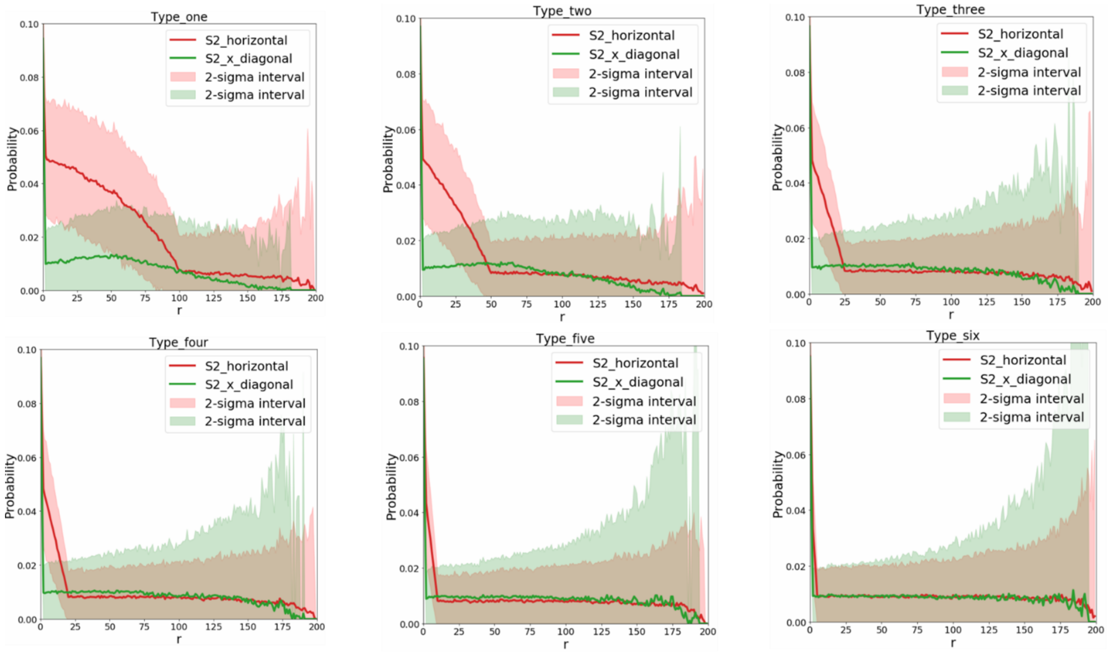


Figure 4.14: S_2 probability as a function of distance r (0 to 200) for the 500 realizations of binary synthetic image of six connectivity types

The red curve indicates the probability in horizontal direction and the green curve indicate the probability in diagonal direction. The red shade and green shade represent the 2-standard deviation of probability at each distance. For each type of the probability response, the probability at distance 0 for either horizontal or diagonal directions is around 0.1, which indicate the proportion for the white phase is around 0.1(around 4000 pixels to 200 x 200 pixels). In the averaged response, the probability starts to drop continuously with small local variations. At each distance below the maximum length of the bar, the red curve stays above the green curve, which indicates the probability in horizontal direction is higher than diagonal direction and suggests that the connectivity in horizontal direction is higher than diagonal direction of white component. Across the six plots, the red curve drops more and more sharply from distance 0 to the maximum length of the white bars, which indicates the short-scale connectivity in horizontal direction decreases from type one to type six. Also, the red curve is getting towards the green curve, which suggests the difference in connectivity between horizontal and diagonal directions is reduced. The two-sigma range for type one is wide below the maximum of length in the bar because the connectivity of these random realizations in type one has great variation. This variation decreases from type one to type six since the dissection and random redistribution operation make the realizations for each type similar to each other gradually. The red and green shade at the tail for all the types are extremely high due to the limited selection of pair pixels at that distance.

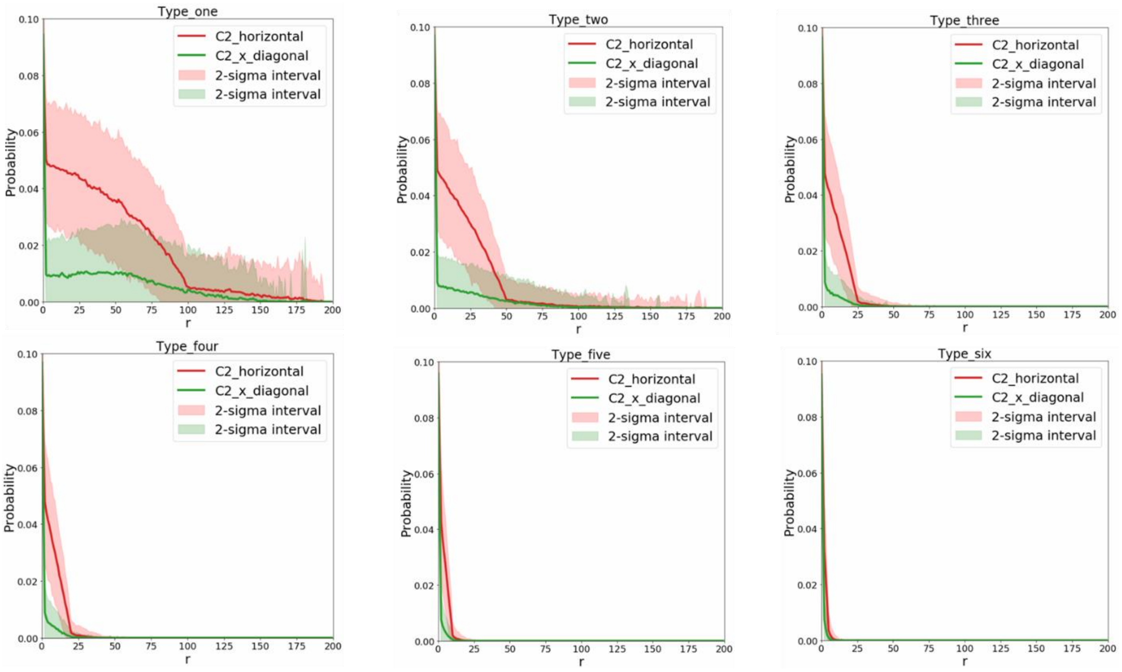


Figure 4.15: C2 probability as a function of distance r (0 to 200) for the 500 realizations of binary synthetic image of six connectivity types

The two-point cluster function is calculated in orthogonal and diagonal directions as well.

The C_2 responses are shown in **Fig 4.15**. The red and green curves are the probability of C_2 response at distance from 0 to 200. Starting from around 0.1 probability at distance 0, the same trend is observed that horizontal connectivity larger than diagonal connectivity for all the six types of images. It can be clearly seen that connectivity in both horizontal and diagonal directions decrease from type one images to type six images. One thing in C_2 results differs from S_2 results is that from a certain distance on, the probability starts to maintain 0 since pair pixels in different cluster does not count.

4.2.2 Fast Marching Method Results for Synthetic Dataset

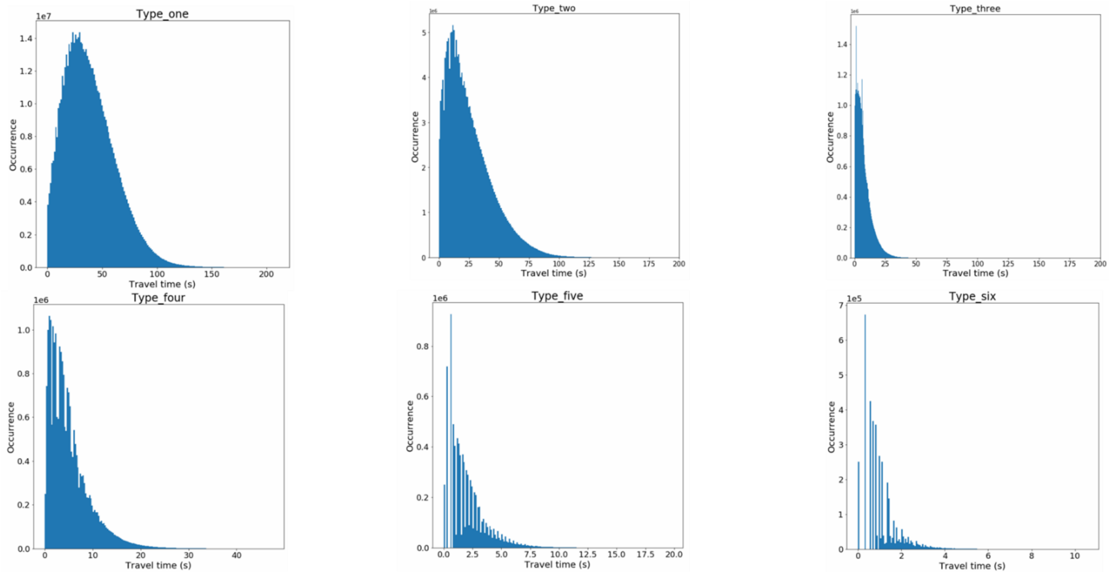


Figure 4.16: Histogram of travel time summarized across 500 images for the six types of images

In the fast-marching process, we first pick 500 random initializations of starting point of source wave, where each of the point is located at pixels of bars in white phase. For setting the travel speed, the speed for wave traveling in white phase and background is set to be 3 m/s and 0 m/s, respectively, where a pixel length represents to 1 m. According to the travel speed, the distance from each pixel to the source point, the travel time is therefore recorded for pixels that the wave can reach. The histogram shown in **Fig 4.16** for each of the type is generated by grouping the time responses of pixels that are reached during the 500 random initializations across all the images in the type. Horizontal axis is the travel time in seconds from source wave to the pixels. Vertical axis is the occurrence at each bin of travel time. From the plots, a left shift of maximum travel time toward original point can be observed from type one to type six, which indicates images in type one has the longest travel distance within a cluster from the source wave to pixels that can be reached. Since the background pixels have no travel speed during fast marching, they block the wave and stop it to transmit to other pixels in white phase, a substantial drop in percentage of pixels being reached should be observed. Assume that all the white pixels are connected, by random picking a location to be the starting point of source wave, all the white pixels could be reached and each of them would have a unique travel time. Thus, the connectivity can be compared by the ratio of number of pixels being reached to the number of pixels that should be reached if they form a single cluster. However, due to the random picking of the source wave, this ratio should be averaged across sufficient number of initializations. In practice, we first gathered the summation of the number of pixels that are reached in each of the 500 initializations and the number of white pixels in each image. Then we obtained the percentage of pixels being reached by taking the average. From type one to type six, the percentages are 0.68, 0.2, 0.037, 0.025, 0.009 and 0.004 respectively. The percentage decreases substantially across the six type images during fast marching, which indicates the connectivity drops significantly.

4.2.3 Cluster Size Distribution Results for Synthetic Dataset

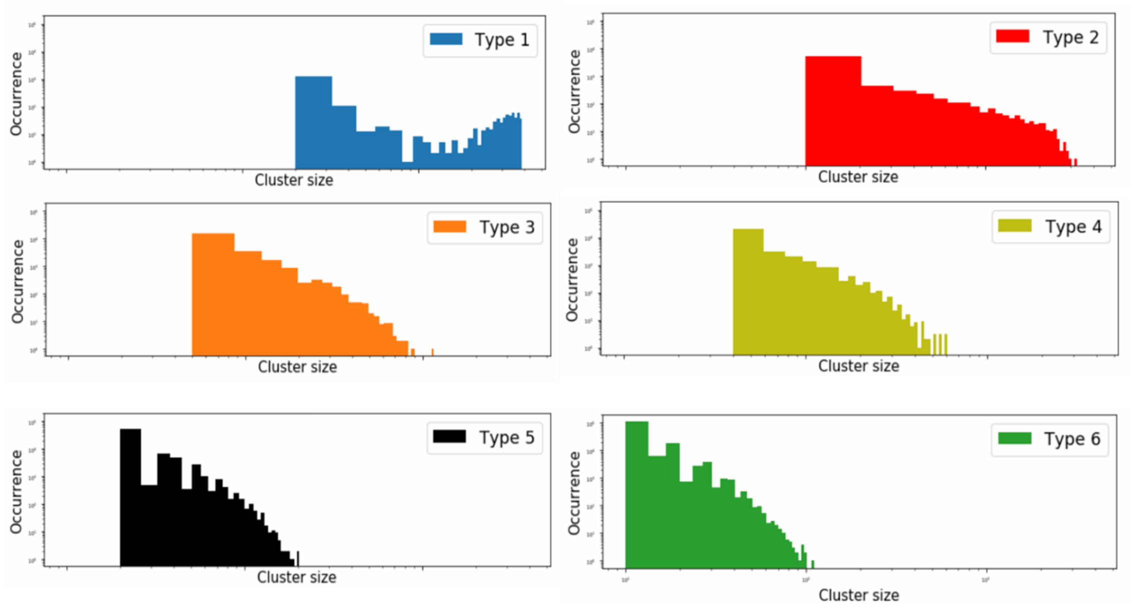


Figure 4.17: Cluster size distribution of 500 images for each of the six types of images

The number of clusters and the size of each cluster are recorded and combined type-wise. The histograms of cluster size distribution for the six types are shown in **Fig 4.17**.

The horizontal axis represent size of clusters in log scale and the vertical axis is the occurrence for each bin of cluster size. For type one images, the chance that a group of several white bars get connected is high so that larger cluster size is observed most often, which results in a rise in the tail of the histogram. By comparison, the histograms shift towards the left can be observed, which indicates the average size of clusters decreases across the six types of images. Thus, the averaged cluster size could be an indicator of connectivity for comparison.

4.2.4 Euler's Number Results for Synthetic Dataset

The result of Euler's number for the six types of images are shown in the **Fig 4.18**.

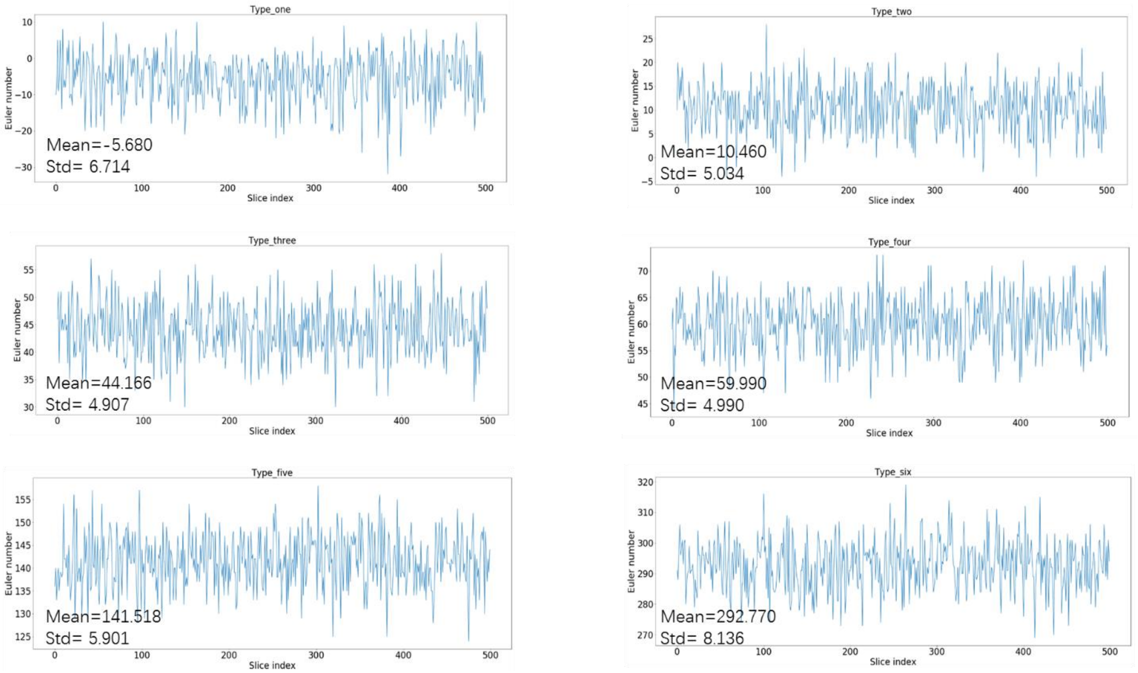


Figure 4.18: Euler's number for each image in the six types

In each plot, the Euler's number is shown for the 500 random initializations. For the type one images, the average Euler's number is -5.68 with standard deviation of 6.714. The average of Euler's number indicates a high connectivity for the type one images. The standard deviation indicates the variations within the type. The averaged Euler's number increases from -5.68 to 292.770, which means that as white bars get dissected continuously, more clusters formed, which result in substantial increases in the minuend such that a decrease in connectivity from type one to type six images can be observed and there is no overlap of the range of Euler's number among the six types.

4.2.5 Results Comparison between Real SEM Images

We conduct connectivity quantification on real SEM images in this section. The images are selected from segmented results in the segmentation part. The slices have four components in it, namely, pores and cracks, organic matter, rock matrix and pyrite.

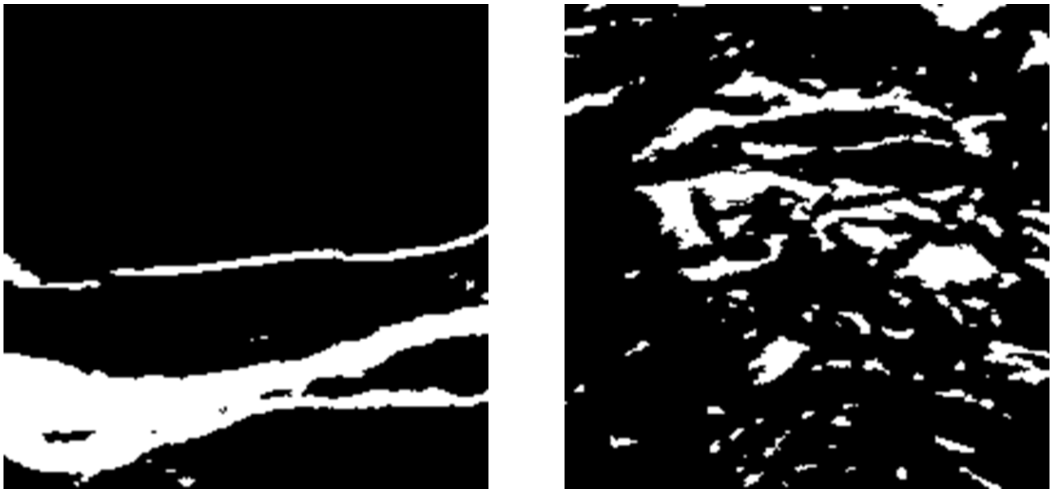


Figure 4.19: Organic matter in the two images shows different connectivity where the connectivity of the first image is substantially higher than the second one.

We first convert the segmented images into binary images such that component of interest is masked as 1 and the rest to be background as 0, where the component of interest represents the component we perform quantification of connectivity of. The two binary images shown in **Fig 4.19** have the same image size of 200 pixel by 200 pixel.

In this study, organic matter and pores and cracks are our components of interest. In the figure, the white phase in the two images represent organic matter and the black represent background. The proportion of the organic matter in the two images are the same, 0.15. Visually we can differentiate that the organic matter in the first image has higher connectivity than the second one. The assumption that the responses from our metrics should be different for these two images and one can tell from the responses which one has higher connectivity is made.

The S_2 and C_2 responses for the two images are shown in **Fig 4.20**.

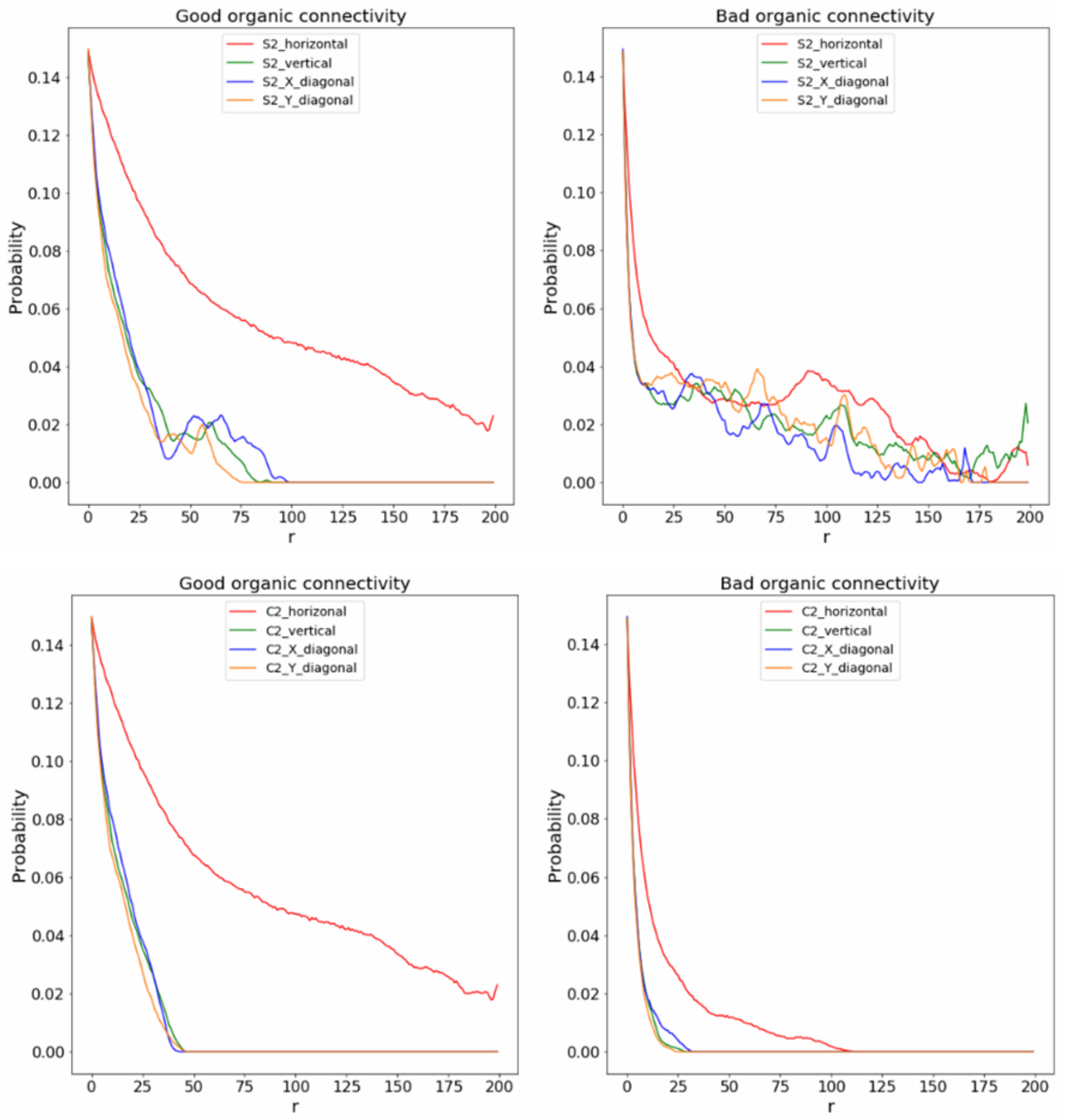


Figure 4.20: Images on top are S_2 response and images at bottom are C_2 response for the two images respectively

Since the assumption that two directions in orthogonal, two directions in diagonal are the same for synthetic data set does not hold true for these two real images, the S_2 and C_2 responses thus are shown in four directions. The red, green, blue and orange lines represent horizontal, vertical, X_diagonal and Y_diagonal directions respectively. It is clear seen that the C_2 probability of the first image in all the four directions drops more gradually at the first several distance than that of the second image, which indicates the connectivity for the first image is higher than the second one. The red line drops more gradually compared to the rest directions, which suggest that the probability at each distance in horizontal direction is higher than that in the rest three directions indicating that the connectivity in horizontal direction are the highest.

Travel time responses are gathered using fast marching method on the two images. The histogram of travel time is shown in **Fig 4.21**.

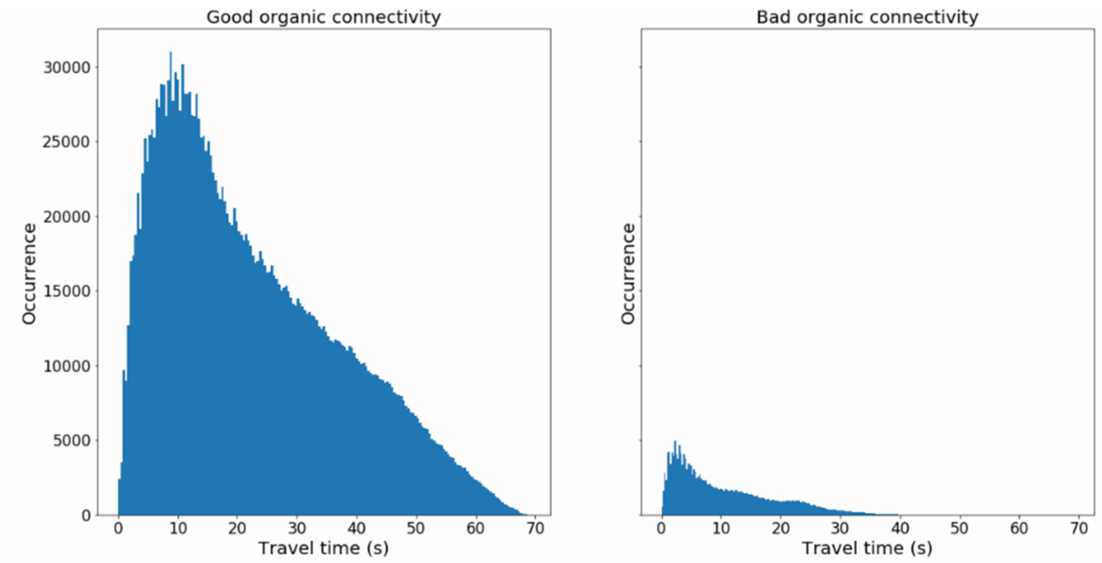


Figure 4.21: Histogram of travel time obtained from fast marching process for the two images respectively

The occurrence at each bin of travel time for the first image is higher than the second one. The mean travel time for the first image is 25s and mean travel time for the second one is 10s. The mean value for the first image is higher than that of the second one indicates that the source wave can reach to pixels far away from it. Also, the percentages of pixels being reached during the fast-marching process are 0.78 and 0.07 respectively. The two observation suggests that the connectivity of the first image is much higher than the second one, which agrees with the conclusion of the visual observation, and S_2 and C_2 responses.

Euler's number for the two images are 6 and 105 respectively, showing that the connectivity for the first image is higher than the second one.

The connectivity of pores and cracks in the study is also being quantified. For a simple demonstration, two binary images shown in **Fig 4.22** have the same image size of 200 pixel by 200 pixel, where the white component represents pores and cracks and the black component represents background. The proportion of white component in the two images are the same, 0.043.

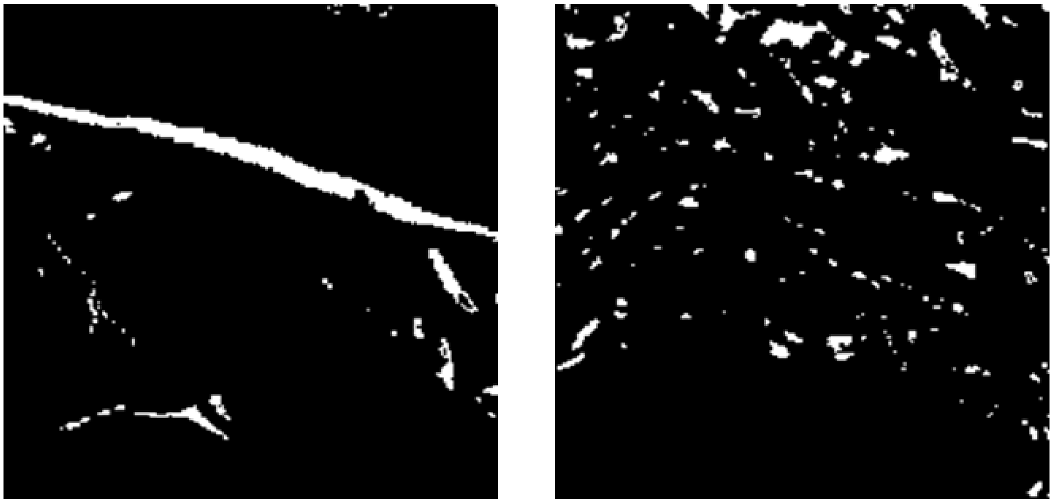


Figure 4.22: Pores and cracks in the two images shows different connectivity where the connectivity of the first one is substantially higher than the second one.

The metrics are directly applied to the two images and responses are shown in **Fig 4.23, 4.24.**

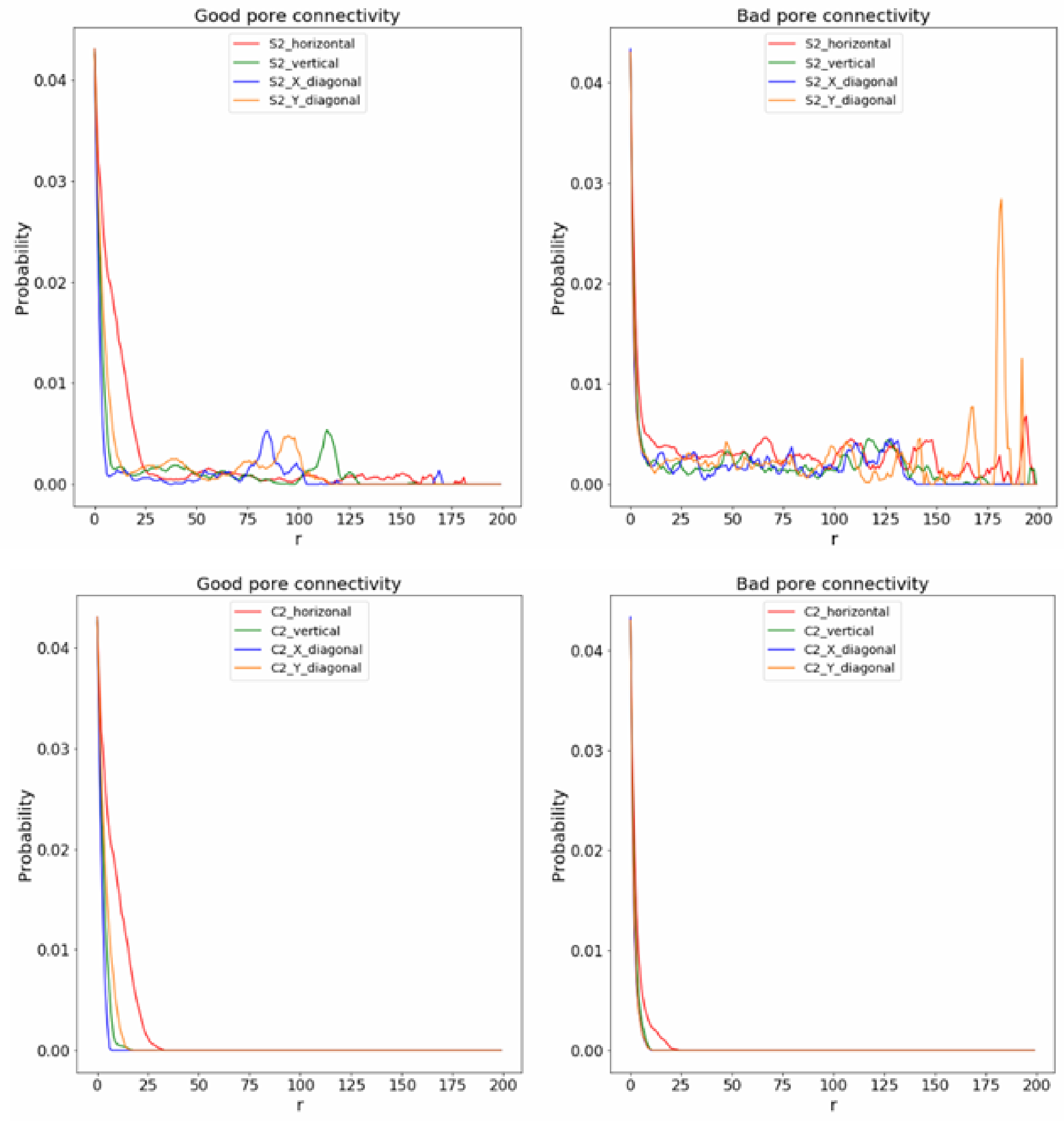


Figure 4.23: Images on top are S_2 response and images at bottom are C_2 response for the two images of pores and cracks, respectively

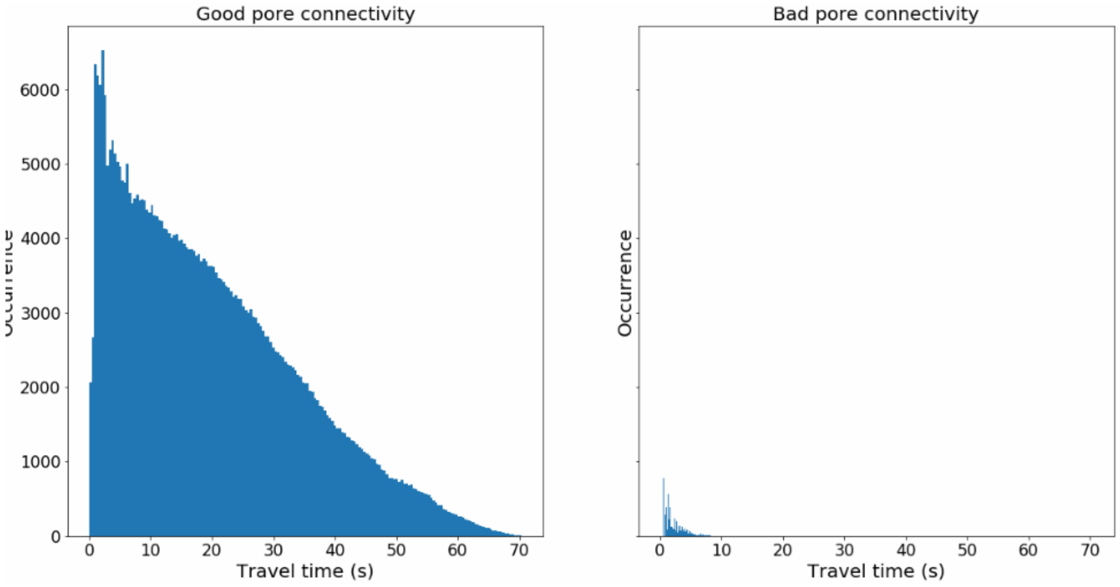


Figure 4.24: Histogram of travel time obtained from fast marching process for the two images of pores and cracks, respectively

Observation of S_2 and C_2 response of the second image that much variations in S_2 response as the distance goes higher compared with 0 probability in C_2 response indicates the clusters are scattered over the image. At each given distance, the probability of C_2 of the first image is higher than that of the second one, which suggests that horizontal connectivity of pores and cracks in the first image is higher than the other one. Based on the histogram of travel time, the average travel time and percentage of pixels being reached for the first image are calculated as 20.069s, 45.55%, whereas 2.065s, 2.13% are obtained for the second image. Euler number are determined to be 45 and 170 for the two images, respectively. All these responses show that connectivity of pores and cracks in the first image is much higher than the second one.

Chapter 5: Conclusions and Limitations

5.1 Conclusions

Machine-learning-assisted segmentation workflow successfully located kerogen/organic, pore/crack, pyrite, and matrix components in SEM images of shale samples. The model was trained on 705, 15000, 17373 and 2074 pixels representing the four components, respectively. The trained method successfully segmented SEM images of size 2058 pixels by 2606 pixels. The model deployment takes an average of 30 seconds on an Intel Xeon CPU E5-1650 v3 @ 3.5GHz, 32GB RAM desktop computer to segment a single SEM image of that size.

Average F1 scores of the segmentation for both inner and transition regions are 0.94, 0.97, 0.8, and 0.83 for (1) pore/crack, (2) organic/kerogen, (3) matrix, and (4) pyrite, respectively. The method is shown to be superior to the threshold-based method, object-based method, and the Fiji segmentation plugin. The segmentation method is demonstrated to be reliable for differentiating pore/crack from organic/kerogen in both inner region and transition zone.

Five different connectivity-quantification metrics, namely two-point statistical function (S_2), two-point cluster function (C_2), cluster size distribution, travel times computed using fast marching method (FMM), and Euler's number, are tested on synthetic dataset of binary images and applied on SEM segmented images. The area under the curve for C_2 are the indicator of connectivity for the four directions. S_2 response serve as the compliment for C_2 function to measure how cluster are distributed. The averaged travel time and the percentage of pixels being reached is used as indicator of connectivity. Euler's number is compared directly for different images. The relationships between the connectivity and the responses of the five connectivity-quantification metrics are determined and validated by statistical analysis on a synthetic dataset of binary images, which contains six types of connectivity from the lowest to the highest. The relationships are directly applied to quantify the connectivity of organic/kerogen and pore/crack components in the SEM images of shale. According to the work on the synthetic dataset, among our connectivity metrics, the best method is Euler’s number because one can quickly access to the connectivity comparison among these types of images by looking at discrete integers. The second-best indicator is the histogram of travel time from the fast-marching method since it contains not only distance but also the information about the full path between connected pixels irrespective to directions. C_2 and S_2 plot are also good indicators because they not only contain information about the magnitude of connectivity, but also the directional and spatial features of the connectivity. The worse method is cluster size distribution since it is hard to describe how the distribution of clusters will lead to the conclusion about difference in connectivity given only small number of images.

5.2 Limitations and future work

For image segmentation, misclassification of pixels still exists in transition zone. Only four components can be identified and segmented accordingly. The annotation process is time consuming because of the manual selection of pixels. For connectivity quantification, the connectivity of components can only be compared by the responses of the metrics, which are indirect indicator. S_2, C_2 and FMM metrics are computational expensive even on a 200 pixel by 200 pixel image. The effect of image size and volume fraction of components on the connectivity are not well understood.

In future work, following tasks need to be accomplished to address existing limitations of our study:

For image segmentation: (1) improve the capability of the method to segment seven components, namely pyrite, kerogen/organic, clay, quartz, organic pore, inorganic pore, and cracks; (2) improve the segmentation performance for the pixels in the transition zone by improving feature extraction and models; (3) apply unsupervised learning and deep learning techniques to improve feature extraction and classification; and (4) more investigation is required to understand the generalization capability of the proposed segmentation method and to compare against existing traditional segmentation methods on images of various types of geomaterials.

For connectivity quantification: (1) the effect of size and volume fraction of components on the connectivity should be further investigated; (2) how the image quality will affect the connectivity quantification (3) find out ways to reduce computation time for S_2, C_2 and fast marching method (4) investigate 3D connectivity

References

[1] H. Sone and M. D. Zoback, "Mechanical properties of shale-gas reservoir rocks—Part 1: Static and dynamic elastic properties and anisotropy," *Geophysics,* vol. 78, no. 5, pp. D381-D392, 2013.

[2] D. Orozco and R. Aguilera, "A Material Balance Equation for Stress-Sensitive Shale Gas Reservoirs Considering the Contribution of Free, Adsorbed and Dissolved Gas," in *SPE/CSUR Unconventional Resources Conference*, 2015: Society of Petroleum Engineers.

[3] L. He, H. Mei, X. Hu, M. Dejam, Z. Kou, and M. Zhang, "Advanced Flowing Material Balance To Determine Original Gas in Place of Shale Gas Considering Adsorption Hysteresis," *SPE Reservoir Evaluation & Engineering,* 2019.

[4] J. E. Johnston and N. I. Christensen, "Seismic anisotropy of shales," *Journal of Geophysical Research: Solid Earth,* vol. 100, no. B4, pp. 5991-6003, 1995.

[5] B. Driskill, J. Walls, J. DeVito, and S. W. Sinclair, "11 Applications of SEM Imaging to Reservoir Characterization in the Eagle Ford Shale, South Texas, USA," 2013.

[6] H. T. Tran, J. D. Jernigen, M. E. Curtis, C. H. Sondergeld, and C. S. Rai, "Investigating Microstructural Heterogeneity in Organic Shale via Large-Scale, High-Resolution SEM Imaging," in *Unconventional Resources Technology Conference, Austin, Texas, 24-26 July 2017*, 2017, pp. 14-25: Society of Exploration Geophysicists, American Association of Petroleum ….

[7] P. Wang *et al.*, "Heterogeneity of intergranular, intraparticle and organic pores in Longmaxi shale in Sichuan Basin, South China: Evidence from SEM digital images and fractal and multifractal geometries," *Marine and Petroleum Geology,* vol. 72, pp. 122-138, 2016.

[8] M. E. Curtis, C. H. Sondergeld, R. J. Ambrose, and C. S. Rai, "Microstructural investigation of gas shales in two and three dimensions using nanometer-scale resolution imagingMicrostructure of Gas Shales," *AAPG bulletin,* vol. 96, no. 4, pp. 665-677, 2012.

[9] P. D. R. Raju and G. Neelima, "Image segmentation by using histogram thresholding," *International Journal of Computer Science Engineering and Technology,* vol. 2, no. 1, pp. 776-779, 2012.

[10] V. Grau, A. Mewes, M. Alcaniz, R. Kikinis, and S. K. Warfield, "Improved watershed transform for medical image segmentation using prior

information," *IEEE transactions on medical imaging,* vol. 23, no. 4, pp. 447-458, 2004.

[11] C. Xu, S. Misra, P. Srinivasan, and S. Ma, "When Petrophysics Meets Big Data: What can Machine Do?," in *SPE Middle East Oil and Gas Show and Conference*, 2019: Society of Petroleum Engineers.

[12] A. Rostami, A. Baghban, A. H. Mohammadi, A. Hemmati-Sarapardeh, and S. Habibzadeh, "Rigorous prognostication of permeability of heterogeneous carbonate oil reservoirs: Smart modeling and correlation development," *Fuel,* vol. 236, pp. 110-123, 2019.

[13] J. He and S. Misra, "Generation of Synthetic Dielectric Dispersion Logs in Organic-Rich Shale Formations Using Neural-Network Models," *Geophysics,* vol. 84, no. 3, pp. 1-46, 2019.

[14] A. Rostami, M. Arabloo, M. Lee, and A. Bahadori, "Applying SVM framework for modeling of CO2 solubility in oil during CO2 flooding," *Fuel,* vol. 214, pp. 73-87, 2018.

[15] H. Li and S. Misra, "Long short-term memory and variational autoencoder with convolutional neural networks for generating NMR T2 distributions," *IEEE Geoscience and Remote Sensing Letters,* vol. 16, no. 2, pp. 192-195, 2019.

[16] H. Li, S. Misra, and J. He, "Neural network modeling of in situ fluid-filled pore size distributions in subsurface shale reservoirs under data constraints," *Neural Computing and Applications,* pp. 1-13, 2019.

[17] H. Li and S. Misra, "Prediction of subsurface NMR T2 distributions in a shale petroleum system using variational autoencoder-based neural networks," *IEEE Geoscience and Remote Sensing Letters,* vol. 14, no. 12, pp. 2395-2397, 2017.

[18] J. He, S. Misra, and H. Li, "Comparative Study of Shallow Learning Models for Generating Compressional and Shear Traveltime Logs," *Petrophysics,* vol. 59, no. 06, pp. 826-840, 2018.

[19] R. Anemone, C. Emerson, and G. Conroy, "Finding fossils in new ways: An artificial neural network approach to predicting the location of productive fossil localities," *Evolutionary Anthropology: Issues, News, and Reviews,* vol. 20, no. 5, pp. 169-180, 2011.

[20] T. Bauer and P. Strauss, "A rule-based image analysis approach for calculating residues and vegetation cover under field conditions," *Catena,* vol. 113, pp. 363-369, 2014.

[21] H. Li, J. He, and S. Misra, "Data-Driven In-Situ Geomechanical Characterization in Shale Reservoirs," in *SPE Annual Technical Conference and Exhibition*, 2018: Society of Petroleum Engineers.

[22] C. Wu, H.-P. Cheng, S. Li, H. Li, and Y. Chen, "ApesNet: A pixel-wise efficient segmentation network for embedded devices," *IET Cyber-Physical Systems: Theory & Applications,* vol. 1, no. 1, pp. 78-85, 2016.

[23] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, 2015, pp. 234-241: Springer.

[24] H. Li and S. Misra, "Assessment of miscible light-hydrocarbon-injection recovery efficiency in Bakken shale formation using wireline-log-derived indices," *Marine and Petroleum Geology,* vol. 89, pp. 585-593, 2018.

[25] S. Shen, W. Sandham, M. Granat, and A. Sterr, "MRI fuzzy segmentation of brain tissue using neighborhood attraction with neural-network optimization," *IEEE transactions on information technology in biomedicine,* vol. 9, no. 3, pp. 459-467, 2005.

[26] S. H. Ong, N. Yeo, K. Lee, Y. Venkatesh, and D. Cao, "Segmentation of color images using a two-stage self-organizing network," *Image and vision computing,* vol. 20, no. 4, pp. 279-289, 2002.

[27] Y. Jiang and Z.-H. Zhou, "SOM ensemble-based image segmentation," *Neural Processing Letters,* vol. 20, no. 3, pp. 171-178, 2004.

[28] D. N. Tripathi, L. A. Hathon, and M. T. Myers, "Exporting Petrophysical Properties of Sandstones From Thin Section Image Analysis," in *SPWLA 59th Annual Logging Symposium*, 2018: Society of Petrophysicists and Well-Log Analysts.

[29] S. Budennyy, A. Pachezhertsev, A. Bukharev, A. Erofeev, D. Mitrushkin, and B. Belozerov, "Image Processing and Machine Learning Approaches for Petrographic Thin Section Analysis," in *SPE Russian Petroleum Technology Conference*, 2017: Society of Petroleum Engineers.

[30] K. Rahimov, A. M. AlSumaiti, H. AlMarzouqi, and M. S. Jouini, "Use of Local Binary Pattern in Texture Classification of Carbonate Rock Micro-CT Images," in *SPE Kingdom of Saudi Arabia Annual Technical Symposium and Exhibition*, 2017: Society of Petroleum Engineers.

[31] P. Asmussen, O. Conrad, A. Günther, M. Kirsch, and U. Riller, "Semi-automatic segmentation of petrographic thin section images using a “seeded-region

growing algorithm" with an application to characterize wheathered subarkose sandstone," *Computers & geosciences,* vol. 83, pp. 89-99, 2015.

[32] Y. Zhao *et al.*, "Study on the Classification and Formation Mechanism of Microscopic Remaining Oil in High Water Cut Stage Based on Machine Learning," in *Abu Dhabi International Petroleum Exhibition & Conference*, 2017: Society of Petroleum Engineers.

[33] R. Narasimha, H. Ouyang, A. Gray, S. W. McLaughlin, and S. Subramaniam, "Automatic joint classification and segmentation of whole cell 3D images," *Pattern Recognition,* vol. 42, no. 6, pp. 1067-1079, 2009.

[34] A. Zaimi, M. Wabartha, V. Herman, P.-L. Antonsanti, C. S. Perone, and J. Cohen-Adad, "AxonDeepSeg: automatic axon and myelin segmentation from microscopy data using convolutional neural networks," *Scientific reports,* vol. 8, no. 1, p. 3816, 2018.

[35] A. Hughes, Z. Liu, M. Raftari, and M. E. Reeves, "A workflow for characterizing nanoparticle monolayers for biosensors: Machine learning on real and artificial SEM images," PeerJ PrePrints2167-9843, 2014.

[36] D. Tang and K. Spikes, "Segmentation of shale SEM images using machine learning," in *SEG Technical Program Expanded Abstracts 2017*: Society of Exploration Geophysicists, 2017, pp. 3898-3902.

[37] J. Hooke, "Coarse sediment connectivity in river channel systems: a conceptual framework and methodology," *Geomorphology,* vol. 56, no. 1-2, pp. 79-94, 2003.

[38] C. Amoros and G. Bornette, "Connectivity and biocomplexity in waterbodies of riverine floodplains," *Freshwater biology,* vol. 47, no. 4, pp. 761-776, 2002.

[39] P. King, "The connectivity and conductivity of overlapping sand bodies," in *North Sea Oil and Gas Reservoirs—II*: Springer, 1990, pp. 353-362.

[40] J. M. Hovadik and D. K. Larue, "Static characterizations of reservoirs: refining the concepts of connectivity and continuity," *Petroleum Geoscience,* vol. 13, no. 3, pp. 195-211, 2007.

[41] P. Renard and D. Allard, "Connectivity metrics for subsurface flow and transport," *Advances in Water Resources,* vol. 51, pp. 168-196, 2013.

[42] A. W. Western, G. Blöschl, and R. B. Grayson, "How well do indicator variograms capture the spatial connectivity of soil moisture?," *Hydrological processes,* vol. 12, no. 12, pp. 1851-1868, 1998.

[43] S. Torquato and G. Stell, "Microstructure of two-phase random media. I. The n-point probability functions," *The Journal of Chemical Physics,* vol. 77, no. 4, pp. 2071-2077, 1982.

[44] Y. Jiao, F. Stillinger, and S. Torquato, "A superior descriptor of random textures and its predictive capacity," *Proceedings of the National Academy of Sciences,* vol. 106, no. 42, pp. 17634-17639, 2009.

[45] C. Yeong and S. Torquato, "Reconstructing random media. II. Three-dimensional media from two-dimensional cuts," *Physical review E,* vol. 58, no. 1, p. 224, 1998.

[46] C. Yeong and S. Torquato, "Reconstructing random media," *Physical Review E,* vol. 57, no. 1, p. 495, 1998.

[47] K. M. Gerke, M. V. Karsanina, R. V. Vasilyev, and D. Mallants, "Improving pattern reconstruction using directional correlation functions," *EPL (Europhysics Letters),* vol. 106, no. 6, p. 66002, 2014.

[48] M. V. Karsanina, K. M. Gerke, E. B. Skvortsova, and D. Mallants, "Universal spatial correlation functions for describing and reconstructing soil microstructure," *PloS one,* vol. 10, no. 5, p. e0126515, 2015.

[49] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *International journal of computer vision,* vol. 59, no. 2, pp. 167-181, 2004.

[50] I. Arganda-Carreras *et al.*, "Trainable Weka Segmentation: a machine learning tool for microscopy pixel classification," *Bioinformatics,* vol. 33, no. 15, pp. 2424-2426, 2017.

[51] S. Narkheda. (2018). *Understanding AUC-ROC Curve*. Available: https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

[52] *Precision-Recall*. Available: https://scikit-learn.org/stable/auto_examples/model_selection/plot_precision_recall.html

Appendix A: Sensitivity of the Segmentation to the Choice of the Wavelet

In order to test the sensitivity of the segmentation to the choice the wavelet, we selected three wavelet families in combination of the rest features to train models and test their performance respectively. The performance is reported in terms of precision, recall and F1 score. Table A1 is the model performance using wavelet Haar of filter length of 2. Table A2 is the model performance using wavelet Dauchies of filter length of 4. Table A3 is the model performance using wavelet Coiflet of filter length of 6. For the performance in inner region, F1 score drops slightly as the filter length goes higher. For the performance in transition zone, F1 score for the matrix and pyrite increase from 0.84, 0.85 to 0.85, 0.88 respectively. In terms of overall performance, weighted average of F1 score shows slightly drop for the wavelet Coiflet of filter length of 6. However, the drop in performance is not significant, and we conclude that the segmentation is not very sensitive to the choice of wavelet given the filter length is less than 6.

Table A1: Performance of Random forest model using wavelet Haar of filter length of 2 (other features unchanged) on the test dataset without noise for the four rock components in the image, where IR and TZ stand for inner region and transition zone

Components	Precision		Recall		F1-score	
	IR	TZ	IR	TZ	IR	TZ
Pore & Crack	1.00	0.93	1.00	0.97	1.00	0.95
Organic & Kerogen	1.00	0.96	1.00	0.99	1.00	0.97
Matrix	1.00	0.79	1.00	0.90	1.00	0.84
Pyrite	1.00	1.00	1.00	0.74	1.00	0.85
Weighted Avg.	1.00	0.92	1.00	0.91	1.00	0.91

Table A2: Performance of Random forest model using wavelet Dauchies of filter length 4 (other features unchanged) on the test dataset without noise for the four rock components in the image, where IR and TZ stand for inner region and transition zone

Components	Precision		Recall		F1-score	
	IR	TZ	IR	TZ	IR	TZ
Pore & Crack	1.00	0.92	0.99	1.00	1.00	0.95
Organic & Kerogen	1.00	0.96	1.00	0.99	1.00	0.97
Matrix	1.00	0. 82	1.00	0.89	1.00	0.85
Pyrite	1.00	1.00	1.00	0.78	1.00	0.88
Weighted Avg.	1.00	0.92	1.00	0.92	1.00	0.92

Table A3: Performance of Random forest model using wavelet Coiflet of filter length 6 (other features unchanged) on the test dataset without noise for the four rock components in the image, where IR and TZ stand for inner region and transition zone

Components	Precision		Recall		F1-score	
	IR	TZ	IR	TZ	IR	TZ
Pore & Crack	1.00	0.92	0.98	0.96	0.99	0.94
Organic & Kerogen	1.00	0.97	1.00	0.98	1.00	0.98
Matrix	0.98	0. 80	1.00	0.90	0.99	0.85
Pyrite	1.00	1.00	1.00	0.78	1.00	0.88
Weighted Avg.	0.99	0.92	0.99	0.91	0.99	0.91

Appendix B: Model Dependency on Image Orientation

In order to determine whether the ML model is independent of image orientation, we tested the performance of the model trained and tested on images with 90 degree and 180-degree rotation from the default orientation, respectively. The performance is reported in terms of precision, recall and F1 score. Table B1 is the model performance of using 90-degree images. Table B2 is the model performance of using 180-degree images. Compared to Table 4-1, the precision, recall and F1 score for each component are almost identical to those without rotation. We conclude that our model is independent of image orientation and segmentation results are reliable.

Table B1: Performance of Random forest model trained and tested on images with 90-degree rotation without noise for the four rock components in the image, where IR and TZ stand for inner region and transition zone

Components	Precision		Recall		F1-score	
	IR	TZ	IR	TZ	IR	TZ
Pore & Crack	1.00	0.93	0.99	0.94	1.00	0.93
Organic & Kerogen	1.00	0.96	1.00	0.99	1.00	0.98
Matrix	0.99	0.79	1.00	0.90	1.00	0.84
Pyrite	1.00	1.00	1.00	0.79	1.00	0.88
Weighted Avg.	1.00	0.92	1.00	0.91	1.00	0.91

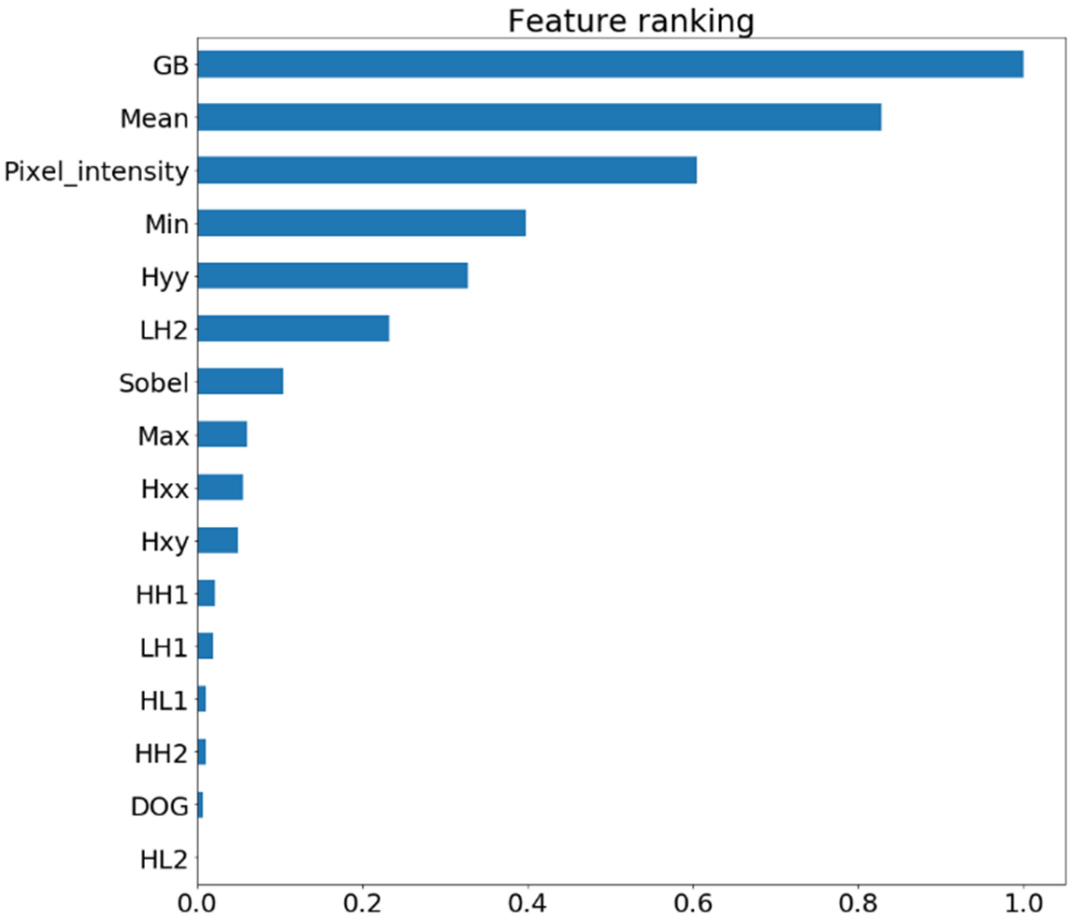


Figure B1: Rank of features in the Random forest model trained on the image with 90-degree rotation

Table B2: Performance of Random forest model trained and tested on images with 180-degree rotation without noise for the four rock components in the image, where IR and TZ stand for inner region and transition zone

Components	Precision		Recall		F1-score	
	IR	TZ	IR	TZ	IR	TZ
Pore & Crack	1.00	0.92	0.99	0.93	0.99	0.92
Organic & Kerogen	1.00	0.96	1.00	0.99	1.00	0.98
Matrix	0.99	0.79	1.00	0.90	0.99	0.84
Pyrite	1.00	1.00	1.00	0.78	1.00	0.88
Weighted Avg.	1.00	0.92	1.00	0.91	1.00	0.91

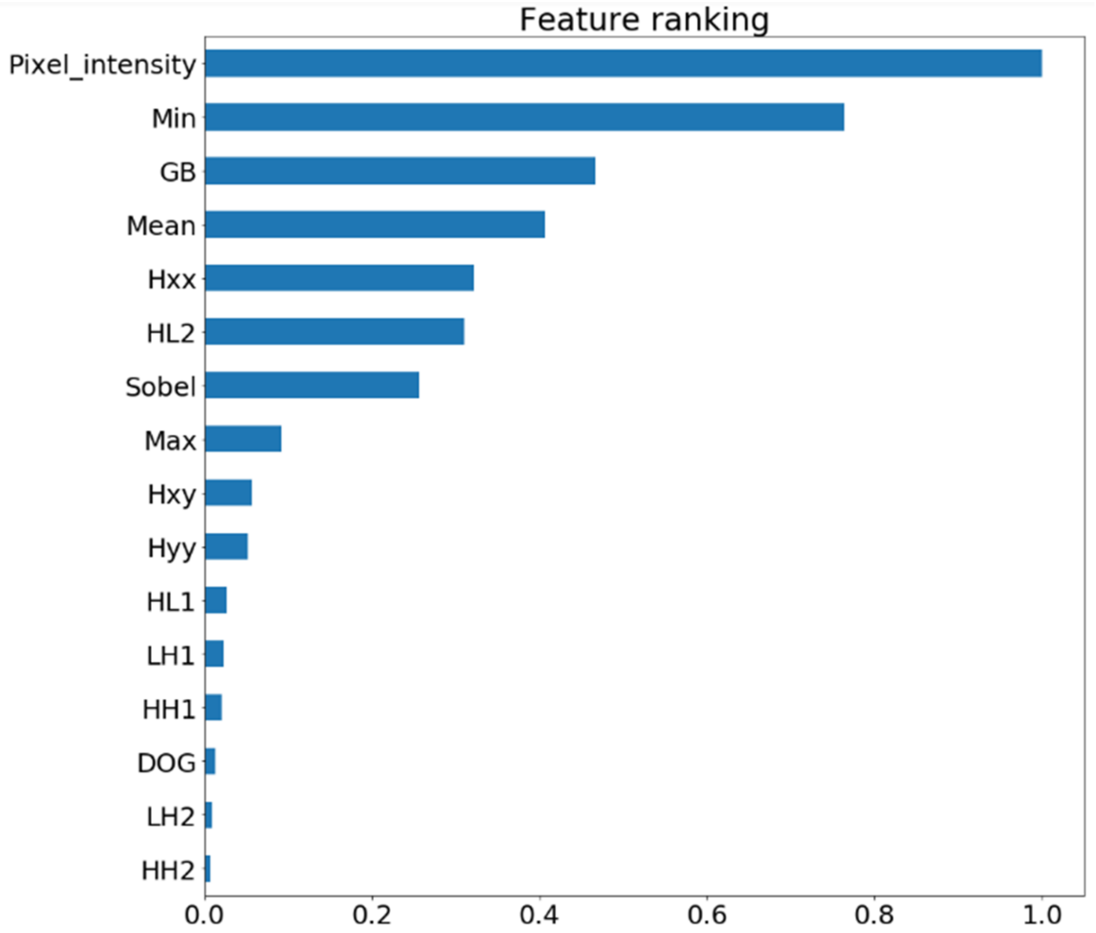


Figure B2: Rank of features in the Random forest model trained on the image with 180-degree rotation

Fig B1 and **Fig B2** show the feature ranking of models trained on image of 90-degree rotation and 180-degree rotation, respectively. The feature ranking shows some variations compared to the original one. For the model trained on images with 90-degree rotation, H_{yy} and LH_2 ranks above H_{xx} and HL_2, which are the top two features when the images are not rotated. The feature ranking between model trained on original image and model trained on 180 degree shows similar results, where features captured in horizontal direction are better ranked than features captured in vertical.

Appendix C: Effect of Image Size on Connectivity Quantification

In order to see the effect of image size on the connectivity quantification, we applied our connectivity metrics on one of synthetic binary image of connectivity type two and the enlarged version of that image, where the original image size is 200 pixel by 200 pixel and the enlarged version is 400 pixel by 400 pixel. **Fig C1** shows the two images used in the study.

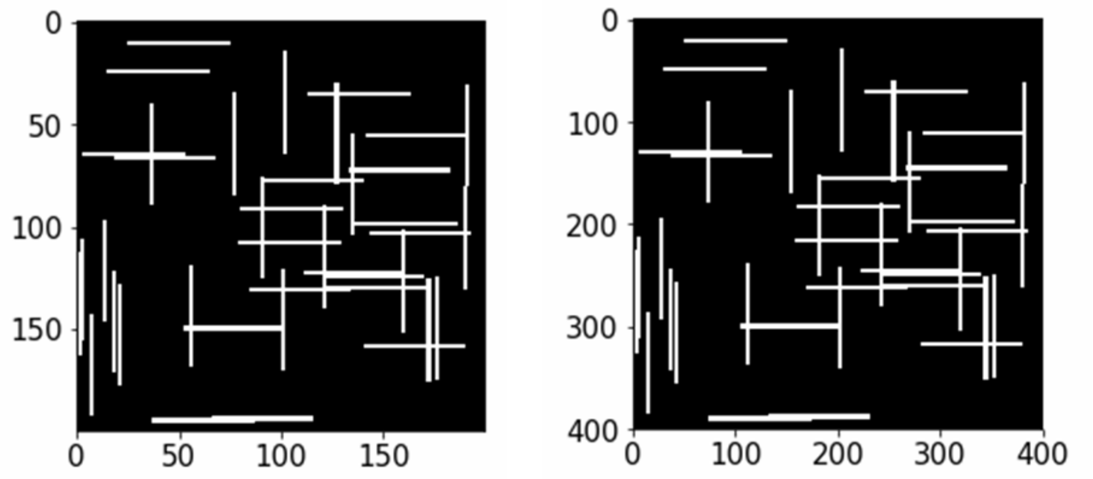


Figure C1: An image from connectivity type two（left）and the enlarged version (right), where the sizes of the left one and the right one are 200 pixel by 200 pixel and 400 pixel by 400 pixel, respectively.

The two images have the same proportion of white phase of approximate 10%.

The S_2 and C_2 responses are shown in **Fig C2**.

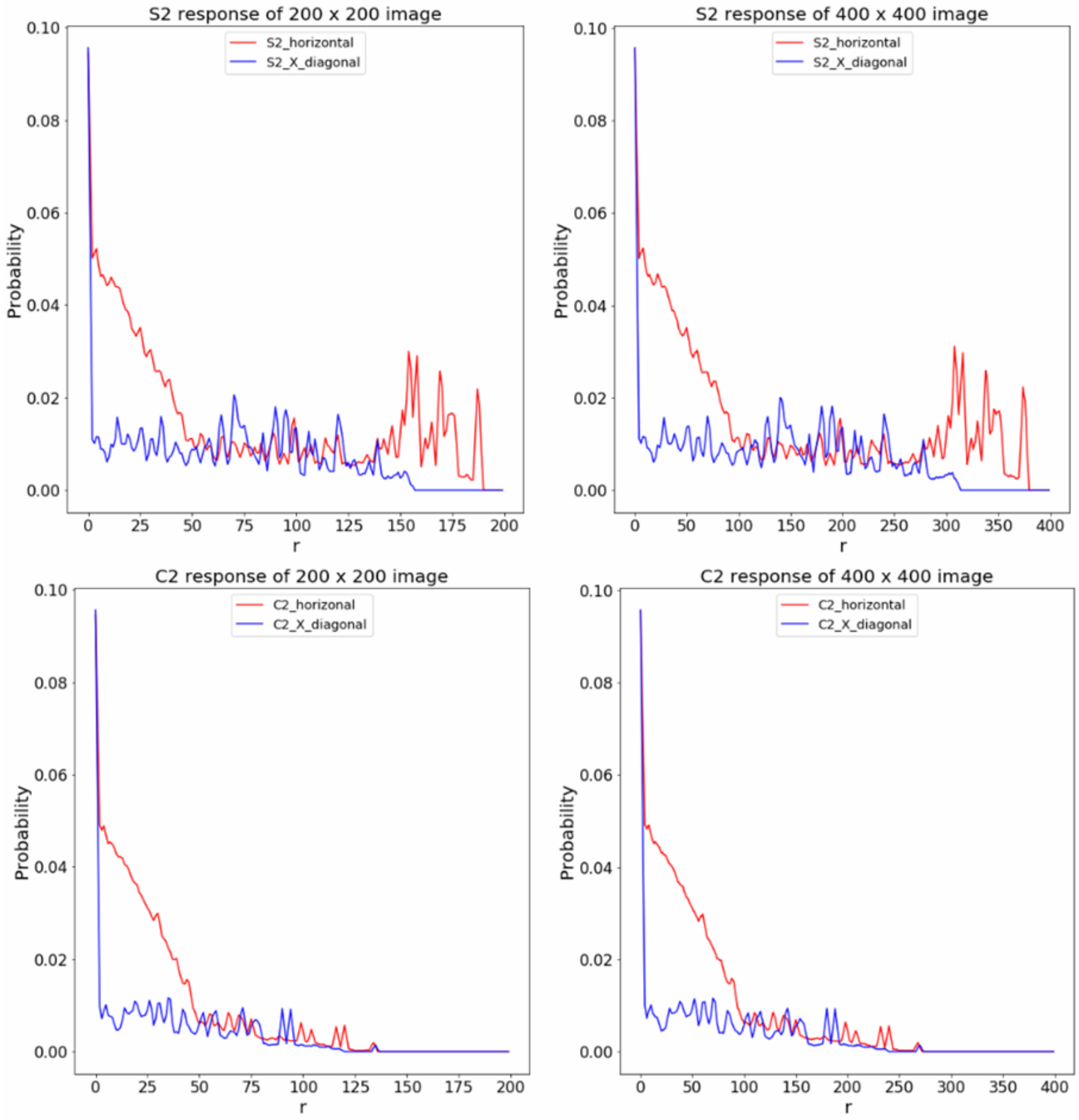


Figure C2: Images on top are S_2 response and images at bottom are C_2 response for the two images, respectively

Red line in the figure represent responses in horizontal direction and the blue line represent responses in X_diagonal direction. The S_2 and C_2 shows the same trend across the length of the images, respectively.

The FMM responses are shown in **Fig C3**.

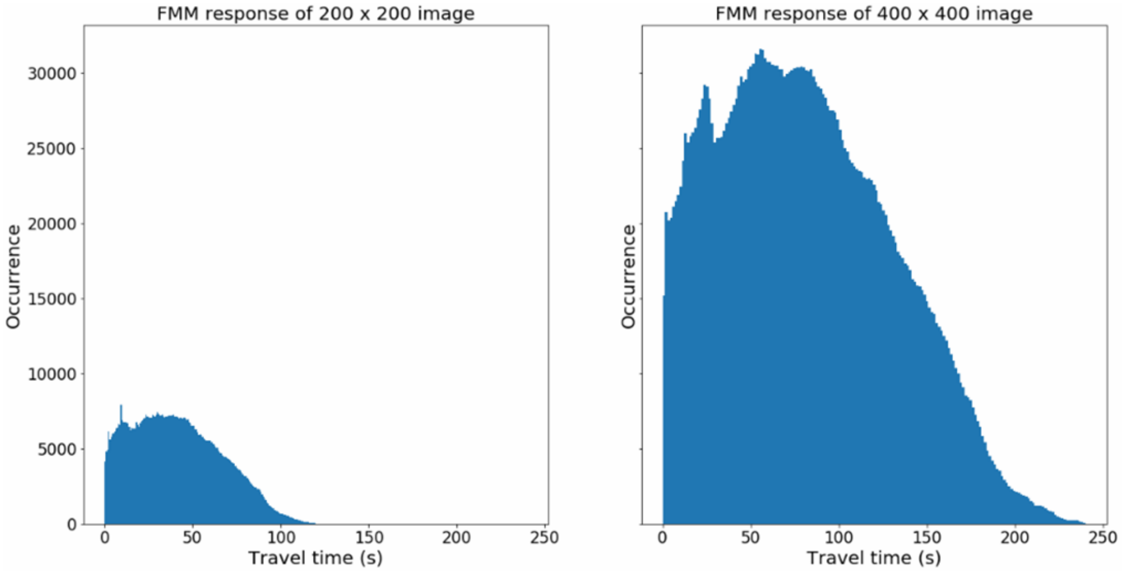


Figure C3: Histogram of travel time obtained from fast marching process for the two images, respectively

The averaged travel time of the first image is 41.4s, whereas that of the second one is approximately a double of the number, which is 81.59s. However, the percentage of pixels being reached are approximately the same for the two images, which are 39.73% and 40.55%, respectively. Finally, Euler numbers for the two images are calculated to be the same, 10.