An object-based approach to differentiate pores and microfractures in petrographic analysis using explainable, supervised machine learning

Issac Sujay Anand John Jayachandran[1], Holly Catherine Gibbs[1], Juan Carlos Laya[1], Yemna Qaiser[2], Talha Khan[2], Mohammed Ishaq Ansari[2], Mohammed Yaqoob Mohammed Shoeb Ansari[2], Mohammed Malyah[2], Nayef Alyafei[3], and thomas Seers[2]

[1]Texas A&M University
[2]Texas A&M University at Qatar
[3]Department of Petroleum Engineering, Texas A&M University at Qatar

September 30, 2023

Study	Thin-sections	Segmented objects	Pore and microfracture classes (Train # — Test #)	Pixel resolution (microns/px)	Size features	Shape features	Features	ML models tested	Training/testing size	Sampling method	Testing accuracies (%)
Ghiasi-Freez et al. (2012)	24	384	Interparticle (134 — 35), intraparticle (39 — 15), oomoldic (62 — 17), biomoldic (25 — 8), vuggy (34 — 15)	-	None	Elongation (aspect ratio), roundness, eccentricity, rectangularity, solidity, equivalent diameter/major axis diameter (they refer to as anisotropy)	6	Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA)	Training = 294, testing = 90	16 largest pores from each sample	66.6 - 100
Borazjani et al. (2016)	39	624	Interparticle, intraparticle, moldic, intercrystalline, vuggy (*class proportions not provided*)	-	Diameter (small, large), perimeter, area, ratio of area to porosity, equivalent diameter, bounding box area, convex hull area	Aspect ratio, roundness, extent, solidity (referred to as stability)	11	Multi-Layer Perceptron (MLP)	Training = 624, testing = 0	16 largest pores from each sample	70 - 100
Mollajan et al. (2016) [*same dataset as Ghiasi-Freez et al., 2012*]	24	384	Interparticle (134 — 35), intraparticle (39 — 15), oomoldic (62 — 17), biomoldic (25 — 8), vuggy (34 — 15)	-	Equivalent diameter	Elongation (aspect ratio), roundness, eccentricity, rectangularity, solidity,	6	Polynomial Support Vector Machine (SVM), k-Nearest Neighbors (kNN), Radial Basis Function Neural Network (RBF-NN), Fusion of all three	Training = 294, testing = 90	16 largest pores from each sample	96.1 - 100
Li et al. (2017) [microCT]	Not specified	Not specified	Vugs and macrofractures	Unclear	Area, perimeter, equivalent circle diameter, effective length, tortuous length, equivalent	Aspect ratio, shape factor, eccentricity	10	Support Vector Machine (SVM)	Not specified	Not specified	Not specified

Dataset: https://doi.org/10.7910/DVN/T2LESU

An object-based approach to differentiate pores and microfractures in petrographic analysis using explainable, supervised machine learning

Issac Sujay Anand John Jayachandran[1,2], Holly Catherine Gibbs[3,4], Juan Carlos Laya[1], Yemna Qaiser[2], Talha Khan[2], Mohammed Ishaq Mohammed Shoeb Ansari[5], Mohammed Yaqoob Ansari[5], Mohammed Malyah[2], Nayef Alyafei[2], Thomas Daniel Seers[1,2]

[1]Department of Geology & Geophysics, Texas A&M University, College Station, TX 77843, USA
[2]Department of Petroleum Engineering, Texas A&M University Qatar, Education City, Doha, Qatar
[3]Department of Biomedical Engineering, Texas A&M University, College Station, TX 77843, USA
[4]Microscopy and Imaging Center, Texas A&M University, College Station, TX 77843, USA
[5]Department of Electrical & Computer Engineering, Texas A&M University at Qatar, Education City, Doha, Qatar

Key Points:

- The first study to propose a binary framing for machine learning driven petrographic pore typing
- Linear and non-linear models perform equally well for idealized microfractures and pores
- We highlight the need for greater scrutiny in AI models for petrographic pore typing

Corresponding author: Issac Sujay Anand John Jayachandran, `sujay92@tamu.edu`

Abstract

Petrographic observations represent a critical aspect of carbonate pore-typing, bridging the gap between the geological framework of a reservoir and its petrophysical behavior. Despite its significance, petrographic pore typing remains a manual endeavor, with the results not easily fitted into quantitative subsurface characterization pipelines. Recent studies have used simplistic pore morphological features within supervised machine learning and deep learning frameworks to automate the petrographic pore-typing process and report strikingly high accuracies in classifying several complex pore types. While supervised learning models are known to be excellent classifiers, most of the literature contains conceptual and technical flaws that raise questions about their validity. Two pore classes that can potentially be separated purely by geometry are microfractures and pores, as they represent intuitive morphological endmembers of the pore system, which should, in theory, maximize the discriminatory utility of simplistic shape features. Also, the use of a binary system as a test case is preferable as supervised machine learning and deep learning models tend to perform strongest for binary classification problems. In the present study, we employed an object-based approach with explainable supervised machine learning to differentiate between open microfractures and open pores viewed in petrographic thin sections. Pores and microfractures were segmented from 18 carbonate thin-sections, sourced from a range of subsurface and outcrop study areas within the USA, and represented numerically by five of the most popular shape features in the geoscientific literature: namely, compactness, aspect ratio, extent, solidity, and formfactor. We used a labeled ground truth dataset containing 400 microfractures and 400 pores to train and evaluate nine of the most widely used linear and non-linear supervised models. All the supervised models performed excellently, with testing accuracies ranging from 89.58 - 90.42%. Notably, the more complex non-linear supervised models did not significantly outperform the simpler linear models, suggesting that the classification of microfractures and pores is a simple, linearly separable problem. In this regard, compactness and aspect ratio were the two most informative features for separating microfractures and pores, with compactness consistently outranking aspect ratio in terms of contribution to the supervised classification. Despite the high accuracies, it was apparent that the labeled dataset of 800 points did not accurately reflect the overall dataset of 20,060 points. While there was excellent separation of the two classes in the labeled data, there was no discernable separation in the global dataset, indicating that the labeled data approximated a complex problem as a simple one. We argue that the high accuracies reported in related studies using similar approaches are more representative of curated datasets than the reality of carbonate pore complexity. We also argue that the simple shape features widely promoted within the geological community may be ineffective towards classifying microfractures and pores and, by extension, higher-order pore types due to their non-unique nature. It is hoped that the results of this study serve as a 'state-of-the-union' for machine learning-assisted quantitative pore typing and lay a foundation for more robust and explainable supervised modeling for pore type classification.

Plain Language Summary

Carbonate pore-typing is a critical task for determining rock types. Petrographic pore typing from thin sections is the most mature form of carbonate pore-typing and is vital in relating the geology of the studied formations to its petrophysical properties. To date, this process has remained manual, bound by human limitations, and difficult to link to quantitative digital reservoir models. Recent research has tried to automate petrographic pore-typing using machine learning and deep learning, claiming very high accuracies. However, there are concerns about these claims due to potential flaws in the methods used. There is potential in using machine learning for binary classification, especially when distinguishing between microfractures and pores, as they are quite distinct in shape. In this study we used an object-based, supervised machine learning approach to differentiate these two classes, using data from 18 carbonate thin sections sourced from

the USA. The data was represented using five popular shape features: namely, compactness, aspect ratio, extent, solidity, and formfactor. We used nine popular linear and non-linear supervised machine learning models. The machine learning models tested had an accuracy of around 90 percent. Interestingly, the more complex non-linear models didn't perform much better than simpler, linear models, suggesting that distinguishing between microfractures and pores might be a straightforward problem. Among the shape features, compactness and aspect ratio proved the most useful in separating the two classes. However, we also report that the labeled dataset used for training the models did not represent the full dataset well, thus indicating that simple shape features cannot accurately capture the complexity of carbonate pore types even at the base binary level. The study concludes that while machine learning is promising for simplistic datasets, we must consider more complex shape features and build larger datasets to develop deep learning models. The hope is that this research will guide future efforts in machine-learning and deep-learning approaches to carbonate pore-type classification.

1 Introduction

Pore classification in carbonate lithologies is a fundamental requirement for subsurface characterization workflows, serving application areas such as carbon capture, utilization, and storage, and hydrocarbon extraction, among others. Critically, carbonate pore-typing serves as the bridge between the geological framework of the subsurface and its petrophysical behavior and is therefore vital to assessing reservoir/aquifer quality (Lønøy, 2006); (Skalinski & Kenter, 2015). Since carbonate pore systems encompass a wide range of scales (nanometric to kilometric scales), holistic pore-typing requires the integration of visual petrographic observations at the thin-section scale with petrophysical data from core plugs and/or well-logs (Skalinski & Kenter, 2015). In this study, we focus on visual petrographic pore-typing, which of the aforementioned data types presents the most direct link to the sedimentological and diagenetic framework of the reservoir and represents the most established modality for pore typing studies (Skalinski & Kenter, 2015; McCreesh et al., 1991).

Visual pore-typing involves user classification of observed pores into types according to popular schema, such as those proposed by Choquette and Pray (1970), Lucia (1983), Lucia (1995), and Lønøy (2006). Presently, visual pore-typing is conducted in a qualitative to semi-quantitative fashion (i.e., via point-counting), a practice that has evolved little since its inception. Barring the inefficiency, subjectivity, and lack of scalability of manual approaches, integrating qualitative / semi-quantitative descriptions into reservoir characterization schemes remains challenging, primarily due to the quantitative nature of the other input data modalities (e.g., well-logs, seismic lines, core plug petrophysical measurements, etc.) (Rabbani et al., 2021).

Recent studies have attempted to automate the process of visual pore-typing, fueled by recent advances in artificial intelligence (AI), and computer vision (CV). These studies attempt to emulate the heuristics employed by geologists when classifying pores by hypothesizing that all pores can be differentiated into their genetic classes purely based on shape. The de facto approach these studies employ is to use supervised machine learning models within an object-based framework, where the segmented pores are represented as objects with size and shape metadata attached (Abedini et al., 2018; Borazjani et al., 2016; Ghiasi-Freez et al., 2012; Mollajan et al., 2016; Z. Wang et al., 2022), as summarized in Table S1 (in supplementary information). Object-based methods are arguably more intuitive for quantitative pore-typing from petrographic images when compared to texture-based methods, as it is easier to recognize geological discontinuities by their size and shape than by their pixel features. Object-based approaches have also become the gold standard in remote sensing studies, collectively referred to as Geographic Object-based Image Analysis (GEOBIA) (Blaschke, 2010).

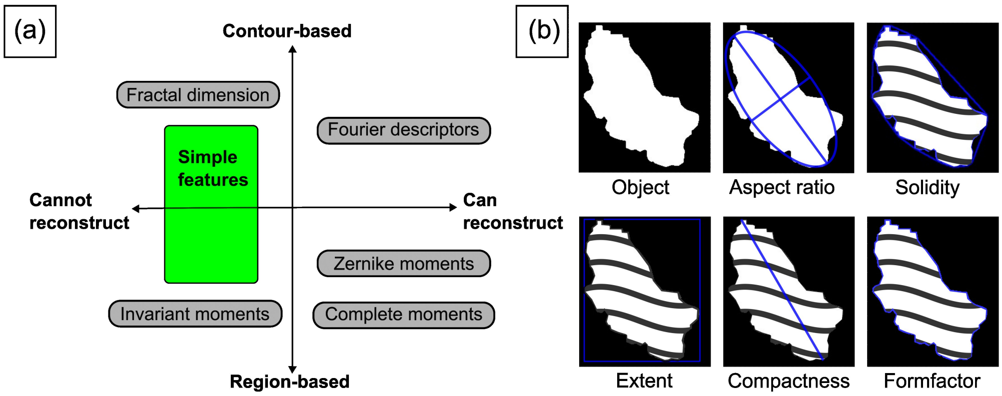

Figure 1. (a) Quadrant of shape features. Modified from (Neal & Russ, 2012). (b) Visual descriptions of the simple shape features used in this study.

Another shared feature amongst most automated pore-typing studies is the use of simple shape features. In the context of pore typing, shape is defined as the geometric features of an object after its location, orientation, and size are removed (Neal & Russ, 2012). Shape features sensitive to location, orientation, and size should be treated with caution (Loncaric, 1998; Neal & Russ, 2012). A useful framework for shape features is the quadrant shown in Fig. 1a. Simple shape features consist of combining size features (such as area, perimeter, maximum axial length of best fit ellipse, etc.) such that the output is a dimensionless ratio (e.g., the ratio of the longest axis to the shortest axis: aspect ratio), in order to remove the influence of scale. While having the benefit of being intuitive and easy to implement, simple shape features also carry the drawback of being non-unique, as several different shapes may have similar feature values (Loncaric, 1998; Neal & Russ, 2012). Conversely, complex shape features, such as Fourier descriptors (harmonic analysis) and moments analyses, while difficult to explain and implement, can reconstruct the original shape of an object and are therefore considered unique to each object (Neal & Russ, 2012). Another critical requirement for shape features is independence (Loncaric, 1998). Each feature must measure unique aspects of the object shape to be informative. If multiple features measure the same property, redundancies occur. Statistical analyses, particularly AI-based methods, can be severely hindered by such redundancies (James et al., 2021; Kuhn et al., 2013).

Relevant literature in the field of quantitative pore typing favor simple shape features to feed ML classifiers (Table S1), reporting testing accuracies well in excess of 90%. These results are remarkable given the complex pore types, such as interparticle, intraparticle, and microfractures (based on the Choquette and Pray (1970) scheme), classified in these studies. Despite these promising results, none of the proposed solutions have widely proliferated within the wider petrographic community attached (Abedini et al., 2018; Borazjani et al., 2016; Ghiasi-Freez et al., 2012; Mollajan et al., 2016; Sharifi, 2022; Z. Wang et al., 2022), with most studies relying upon conventional manual interpretation. This lack of uptake may, in part, be related to a general mistrust in the ostensibly optimistic results published, especially when considering that pore morphology is not the only determining factor when assigning pore types via conventional (i.e., qualitative) means.

Notably, there are deficiencies in four key areas within the literature: (1) the use of natively binary classifiers for multi-class problems, (2) the imbalanced and/or diminutive nature of the input datasets, (3) the lack of robust benchmarking, and (4) the misappropriation of deep learning. Classifiers that are natively binary (esp., Support Vector Machines: SVM) have been employed to classify several different pore types (Mollajan et al., 2016; Sharifi, 2022). For context, binary classifiers can be extended to multi-class problems by condensing them into a series of binary classification problems, typically using a one-versus-all (OVA) or one- versus -the-rest approach (Bishop, 2006; Galar et al., 2011; Mollajan et al., 2016). These approaches are conceptually problematic as the decision boundaries from several binary classifiers are known to create ambiguous regions within the feature space, which can result in the same object being classified as different classes in different iterations (Bishop, 2006). Another inherent flaw is that models are trained on imbalanced data, as the class in focus will typically be diminutive compared to the other classes combined. Notably, such class imbalances are well-known to decrease model performance (Bishop, 2006; Galar et al., 2011; Chawla et al., 2004; He & Garcia, 2009; Sun et al., 2009). Furthermore, as the 'other' classes are typically merged for each classifier, any relationships or dependencies between classes may be ignored. In addition, since the number of binary classifiers will increase linearly with the number of output classes, computational cost, and scalability can rapidly become limiting factors (Bishop, 2006; Galar et al., 2011)

Supervised ML models are particularly sensitive to the nature of the labeled data. Most related studies are opaque on their sampling protocols, which raises questions as to whether the data was properly curated (Table S1) (e.g., Abedini et al., 2018; Borazjani et al., 2016; Ghiasi-Freez et al., 2012; Mollajan et al., 2016; Sharifi, 2022; Z. Wang et al., 2022). There are several indicators within the literature that point towards improper dataset curation; firstly, the aforementioned studies contain severe class imbalances in their training and testing data, which tends to give rise to model instabilities and poor performance (Bishop, 2006; Galar et al., 2011; Chawla et al., 2004; He & Garcia, 2009; Sun et al., 2009) / (Table S1). Secondly, their sample sizes are limited, even going as low as five objects per class within some studies (Table S1) (Abedini et al., 2018; Ghiasi-Freez et al., 2012; Mollajan et al., 2016). The sample sizes are far too insufficient for the complexity pursued to produce robust models (Sun et al., 2009). Finally, several pore types classified are not perceived by shape alone but by the spatial context of skeletal, depositional, and diagenetic components. For example, pore types such as vugs, molds, intraparticle, interparticle, and intercrystalline pores cannot be differentiated by shape but by examining their local neighborhoods. This raises questions about the subjectivity of the labelling process and, therefore, the validity of the training and testing dataset.

There is also a noticeable lack of model benchmarking within the related literature, with supervised machine learning models being arbitrarily chosen to perform a given classification task (Table S1). In addition, several studies embrace deep learning (DL) models, despite the 'excellent' performance of ML models (Abedini et al., 2018; Borazjani et al., 2016; Mollajan et al., 2016; Sharifi, 2022; Ansari, Abdalla, et al., 2022). The associated datasets do not meet the typical class balance and quantity requirements to ensure DL model generalizability. Also, these studies do not provide metrics such as validation-loss curves to provide assurances on the model's accuracy and stability.

A more equitable approach would be to condense pore-typing into a binary classification problem, such as distinguishing between microfractures and pores, as they represent visually distinct endmembers in morphology and are distinct in the mode of genesis. This framing plays to the strength of most supervised ML classifiers as some were designed to be binary classifiers (Multiple Logistic Regression and SVM, among others), and single decision boundaries between two classes are far simpler to construct for any model (Bishop, 2006; Galar et al., 2011; James et al., 2021; Kuhn et al., 2013; Kuhn & Silge, 2022; Ansari, Yang, et al., 2022). In addition, binary classifications also enable additional model performance metrics such as the Receiver Operating Characteristic (ROC) curves (James et al., 2021; Kuhn et al., 2013; Kuhn & Silge, 2022). It is important to note that while performance metrics such as ROC curves can be extended for multi-class problems, it is far more challenging to implement and interpret. Once the end members have been satisfactorily classified and decision boundaries established, it should be possible to analyze intra-class datasets to make finer distinctions between pore and microfracture types. Additionally, due to the ease of recognizing microfractures from pores, the quality of the labeling data would be significantly higher than dividing the pores into genetic types.

Only two studies have employed the binary approach within macrofractures in micro-CT models (Li et al., 2017; Singh et al., 2021), and one in the case of microfractures (Z. Wang et al., 2022). Li et al. (2017) utilized an SVM to separate macrofractures from vugs using simple shape features, reporting an accuracy of 100%. However, the authors did not offer sufficient details on the modeling procedure, and from the images provided, the macrofractures appeared simplistic (short and straight). Singh et al. (2021) demonstrated excellent segmentation of macrofractures and pores (with classification accuracies above 96%) using a projection-based clustering approach comprised of Principal Components Analysis (PCA) and k-means clustering. However, the proposed method cannot be scaled down to microfractures, given that size itself served as a major discriminator between the macrofractures and pores. Z. Wang et al. (2022) reported near-perfect accuracies, nullifying the challenge of classifying microfractures and pores. However, their classification methodology was not described in detail, and the objects sampled for classification were heavily curated and too few to be considered representative.

We propose that employing simple shape features for object classification within a supervised machine-learning framework can accurately determine microfractures from pores. In this work, we pose two questions: firstly, how accurately can supervised models classify microfractures and pores using only simple shape features? We posit that the combination of simple shape features within a supervised ML framework should accurately capture the shapes of microfracture and pores, given that these shapes represent morphological endmembers. We eschewed unsupervised models for this study as supervised models are known to be substantially stronger. However, we did include two clustering algorithms (K-means and DBSCAN) on the global dataset as a reference against the supervised models results (Fig. S4). Secondly, provided a sufficiently high accuracy from the supervised classifiers, we pose the question: what are the most informative simple shape features for differentiating microfractures and pores? We hypothesize that aspect ratio is the most important shape feature as elongation is the primary and most intuitive discriminator between the two classes.

The hypotheses in this study were tested on 18 petrographic plane-polarized light scans of complete thin sections. The provenance of the microfractures is not considered in this study as it is irrelevant to the tested hypotheses. We notify the reader, given the small size of the dataset, that the results of this study are meant to be explanatory and should not be considered as the most accurate models available. It is intended that the results of this study will serve as a substrate for the development of highly accurate classifiers in future work. More importantly, the study was designed to address the methodological deficiencies of the related literature in terms of data handling and supervised ML modeling as per the guidelines provided by Artrith et al. (2021) and Greener et al. (2022).

Finally, we chose not to pursue DL in this study for the following reasons: firstly, we have not fully realized the potential of ML within geo-images, and secondly, the black box nature of DL means that we replace human subjectivity with machine subjectivity, limiting the ability to draw translatable insights from any resulting classification. Finally, in similitude to many geoscientific applications, difficulties in procuring sufficient training and test data make DL impracticable for the present study. To our knowledge, this study represents the only openly available dataset solely dedicated to microfractures and pores of carbonate thin sections within the geosciences.

2 Methods

2.1 Dataset

We selected eighteen images for this study, sourced from a repository of plane-polarized light scans of carbonate thin sections at Texas A&M University, College Station. The thin sections were scanned whole using the Nikon CoolScan 8000 film scanner at a resolution of 6.35 microns/pixel. The thin sections were sourced from a wide variety of outcrops and subsurface cores. A key criterion for selection was the presence of sufficient open-mode microfractures and pores. Healed microfractures (microveins) were ignored as they require a different form of segmentation and are not within the scope of this study. Eleven of the thin sections were half-stained with Alizarin red and seven thin sections were unstained. The staining, however, did not affect the pore segmentation as all the thin sections were impregnated with blue epoxy. The list of thin sections used and associated metadata is provided in the dataset in the GitHub repository of the study.

2.2 Image processing and segmentation

2.2.1 Pre-processing

A schematic diagram for the entire image processing and machine learning pipeline is provided in Fig. 2. For brevity, only the pertinent information is provided in the text, with the finer details of each stage provided in the Supplementary Information. The edges of all images were cropped prior to pre-processing to remove the blank slide edges. The images were of sufficient quality that pre-processing only required minimal denoising and sharpening. For denoising, the non-local means filter was applied using the 'Non-local means denoising' plugin from the Biomedgroup library in Fiji (Darbon et al., 2008). The non-local means filter was chosen for its excellent edge-preserving capabilities (Buades et al., 2011). An unsharp mask filter was used to restore the sharpness after denoising, using the in-built tool within Fiji, tuned according to each image. The images post-denoising and post-sharpening are included as part of the dataset attached in the supplementary information.

2.2.2 Segmentation

The segmentation of the blue-epoxy-filled pores from thin sections only required thresholding in the HSB (Hue-Saturation-Brightness) color space. However, the low resolution of the available thin-section scans presented complications for the segmentation of microfractures. Microfractures that appeared visually continuous tended to be fragmented into several smaller segments after thresholding in the HSB space despite extensive tuning of the thresholding parameters (Fig. S1). To increase the microfracture connectivity, an independent segmentation was performed in the CIELAB color space, which is a device-independent 3D color space that accurately maps all perceivable colors, thus enabling comparison. The CIELAB segmented image was combined with the original HSB segmented image after post-processing both images. While there was a notable increase in the connectivity of several microfractures (examples shown in Fig. S2), several microfractures were still heavily fragmented. Moreover, microporous matrix zones and microporous grains were segmented as macropores as a byproduct of the aggressive segmentation strategy. The sheer number of microporous zones rendered masking impracticable. For this study, they were approximated as pores, which is reasonable given the similarities in terms of shape for both pore types. Finally, any compromised image regions (e.g., scratch marks or air bubbles) were masked manually.

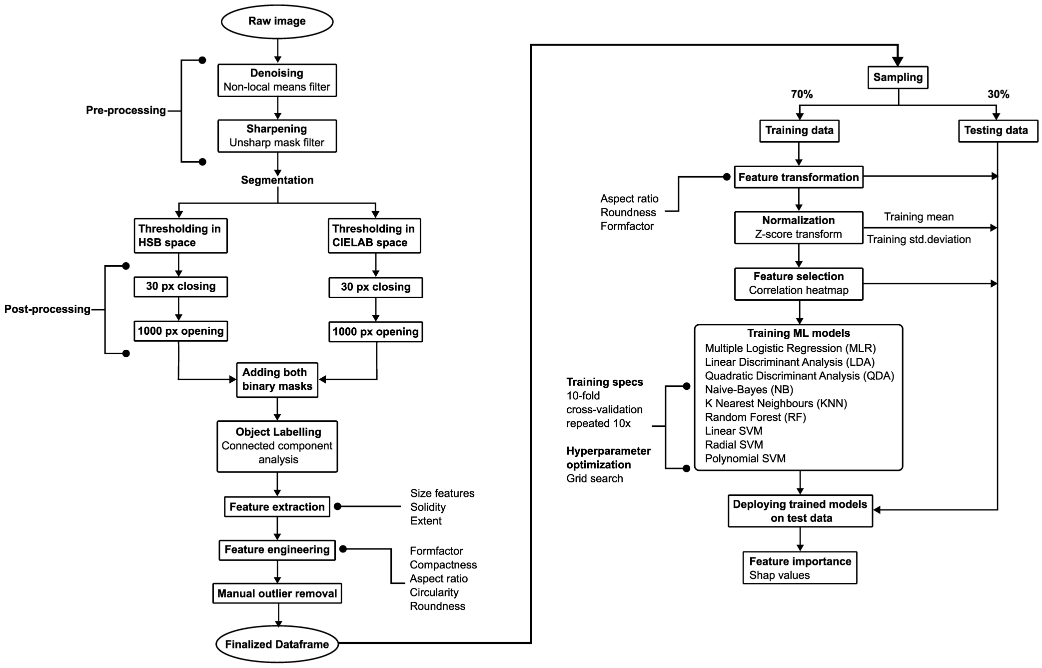

Figure 2. Flowchart of the digital image analysis and supervised modelling workflow.

2.2.3 Post-processing

The post-processing pipeline was conducted on both the HSB and LAB binary masks in parallel (Fig. 2). Binary masks from both color spaces had smaller pores that were poorly resolved, whereby the perimeter of these objects becomes pixelated and/or suffers from partial area effects. As a consequence, the true shape of the pore is lost, and any downstream analysis will be flawed. A workaround is to visually estimate the smallest pore size that is adequately resolved and cull all objects below this threshold. Some studies, particularly in the SEM and Liquid Metal Injection (LMI) domain, refer to the smallest pore that is adequately resolved as the practical pore resolution (PPR) (Hemes et al., 2015) . In this study, we visually estimated that the smallest pore size that was adequately resolved was 30 pixels in area (equivalent to pores of 190.5 microns). A morphological closing operation using a 4-connect was applied using the Gray Scale Attribute Filter tool in the MorphoLibJ plugin in Fiji (Legland et al., 2016), with the conservative 4-connectivity protocol used to prevent microfractures from being removed. It must be noted that, given the relatively poor pixel resolution, the smallest pore size chosen is atypically aggressive, as even at this size, discretization effects are visible in several pores. This aggressive choice was warranted to preserve the microfractures since they were limited in quantity throughout the dataset. Additionally, several of the larger pores had floating objects within them (particles/air bubbles). These were removed from the objects via a morphological opening of 1000 pixels using the Gray Scale Attribute Filter tool in the MorphoLibJ plugin in Fiji.

2.3 Labelling, Feature Extraction and Feature Engineering

The binary masks were imported into Python for labeling and feature extraction. The 'Connected components' function with 8-connect from the OpenCV library was used to label the microfractures and pores. The 'regionprops' module from the Sci-kit image library was used to extract the size and shape features of each object (Table 1). Shape features unavailable in the regionprop module but deemed necessary based on the literature were calculated from the measured size metrics. We note here that eccentricity was discarded, despite its popularity as an elongation metric in the literature, as its distribution was extremely right skewed even after Box-Cox transformations. Representations of the selected shape features are shown in Fig. 1b.

Table 1: Feature Table

Feature	**Equation**	**Definition**	**Shape aspect measured**	**Selected**
Area	NA	The number of pixels of the object	None	No
Filled area	NA	Number of pixels in the object with holes filled	None	No
Convex area	NA	Number of pixels in the convex hull of object	None	No
Perimeter	NA	The number of contour pixels	None	No
Crofton perimeter	NA	Perimeter of object approximated by Crofton formula in 4 directions	None	No
Major axis length	Normalized second central moments	The major axis of the best fitting ellipse	None	No
Minor axis length	Normalized second central moments	The minor axis of the best fitting ellipse	None	No
Equivalent diameter	NA	Diameter of the circle with equal area	None	No
Max feret diameter	NA	Maximum caliper length of object	None	No
Solidity	$\frac{\text{area}}{\text{area of convex hull}}$	Area of the object relative to its convex hull	Convexity	Yes
Extent	$\frac{\text{area}}{\text{area of bounding box}}$	Area of the object relative to its rigid bounding box	Complexity	Yes
Aspect Ratio	$\frac{\text{major axis length}}{\text{minor axis length}}$	Ratio of the major axis to minor axis	Elongation	Yes
Compactness	$\frac{\sqrt{4\times\text{area}/\pi}}{\text{feret diameter max}}$	The ratio of the object area to its maximum Feret diameter	Elongation/circularity	Yes
Formfactor	$\frac{4\times\pi\times\text{area}}{(\text{perimeter crofton})^2}$	Area- and contour-based circularity of the object	Circularity	Yes

Continued on next page

Table 1 – continued from previous page

Feature	**Equation**	**Definition**	**Shape aspect measured**	**Selected**
Eccentricity	$\frac{\text{Distance from Focus}}{\text{Distance from Directrix}}$	Measure of the ellipticity of an object	Elongation/circularity	No
Circularity	$\frac{\text{equivalent diameter}}{\text{perimeter crofton}}$	Outline-based circularity of the object	Circularity	No
Roundness	$\frac{4\times\text{area}}{\pi\times(\text{feret diameter max})^2}$	Area-based circularity of the object	Circularity	No

2.4 Statistical Analysis of the Extracted Features

2.4.1 Outlier Detection

Identifying outliers is a pre-requisite for building machine learning models, as they can hinder model performance and result in convergence to local minima. We omitted automated outlier detection methods (e.g., Tukey's boxplot) due to the aggressive selection criteria such approaches employ. Aggressively removing a large chunk of true objects may improve model accuracy at the cost of generalizability, as the model will overfit to a heavily sanitized training dataset. Consequently, we employed a manual approach, whereby data points that were ten standard deviations from the mean of both size and shape features were visually corroborated with their corresponding thin-section image before being classified as outliers. This manual approach ensured that only the most prominent outliers per image were removed (2-5 per image), thus preserving the potential generalizability of the models. The total number of data points used for modeling was 20,060 after discarding outliers.

2.4.2 Sampling, Primary Labeling and Secondary Labeling

We applied different strategies to sample pores and microfractures, dictated by the limited number of microfractures in the images. Sampling for the pores was performed randomly, while microfractures were sampled manually. 400 microfractures and 400 pores were selected as the labeled dataset. The design of the sampling protocol was intended to maximize the quality of the ground truth. For 100 pores, sampling was performed with pore area greater than 100 pixels to ensure the larger pores were represented in the training and testing sets, given the strong skew towards smaller pores. Open gashes associated with microstylolites were avoided altogether, as these are discontinuities principally formed by pressure solution rather than brittle deformation. Moreover, open gashes were rarely observed in the dataset, and their omission is not expected to impact the results significantly.

To supplement the primary labels of 'pore' and 'microfracture', secondary labels were added to each sampled object pertaining to the type of pore or microfracture. Four types of microfracture were delineated by morphology based on the samples in this study: namely, straight, curvilinear, curved, and branching. These sub-categories were based on visual appearance and not on any established scheme. While labeling microfractures as straight and branching was relatively intuitive, the difference between curvilinear and curved was more subtle. Microfractures that were dominantly linear with negligible deviations were judged as curvilinear, whereas if there were major deviations in their trace morphology, they were classified as curved. Examples of these four types are shown in Fig. 3a. It should be noted that branching microfractures can be further subdivided into further shape-based categories (T-type / X-type, e.g., (Seers & Hodgetts, 2016)), though for parsimony, we avoided such higher-order classes in the present study. Conversely, pore types were defined by origin rather than morphology, namely vug, intercrystalline, intraparticle, and channel, as per the Choquette and Pray (1970). Vug was used as a catchall term applied to group relatively equant pores with evidence of genesis through dissolution and those with ambiguous origin. Intercrystalline pores were those housed within incompletely cemented spaces. Channels posed an interesting conundrum as they originated from microfractures but evolved into pores. However, apart from one sample, channels were rarely observed in the dataset and, therefore, poorly represented. We also point out that interparticle pores were rare in the dataset and, hence, were not represented during the random sampling. The inclusion of sufficient channels and interparticle pores in the training data should be a target for future work.

2.5 Supervised Machine Learning Pipeline

2.5.1 Training-Testing Split

The labeled dataset was split into 70% training and 30% testing subsets in a randomly stratified manner, keeping the proportions of pores and microfractures equal within both sets. This split resulted in 280 microfractures and pores in the training set and 120 microfractures and pores in the testing set. The training-testing split was performed prior to the subsequent data processing to prevent data leakage.

2.5.2 Feature Transformation

All the shape features within the training data exhibited varying degrees of non-normality, with compactness and extent containing visible bimodality, and roundness, aspect ratio, and formfactor showing a degree of right skew. These right-skewed feature sets were log-transformed to balance their data range, mitigating data paucity and potentially increasing model accuracies. We emphasize that the transformation approach was not designed to satisfy the assumption of multivariate normality by parametric models, such as multiple logistic regression (MLR), linear discriminant analysis (LDA), and quadratic discriminant analysis (QDA). The fact that several of the features, post-transformation, were significantly bimodal precludes the possibility of forcibly converting them into normal distributions. Moreover, Graf et al. (2022) showed that LDA is ostensibly robust against lognormal skewed and bimodal distributions, thus indicating that the assumption of normality is not critical. Post-transformation, the features in the training and testing data were centered and scaled to ensure comparability between the features. We note that all features in the testing data were centered and scaled using the mean and standard deviation derived from the training data.

2.5.3 Feature Selection

Feature selection was entirely supervised based on a priori knowledge of the features and their correlations. As discussed above, feeding redundant features into ML models can undermine each feature's true impact and cause model instabilities: a problem known as multicollinearity (James et al., 2021; Kuhn et al., 2013). Furthermore, reducing the number of features decreases the possibility of sparse distributions in feature space, often referred to as the 'curse of dimensionality' (Kuhn & Johnson, 2019). We expected high correlations between the features as each was derived from the same pool of size features. Features with Pearson's correlation coefficient r^2 values exceeding 0.95 were candidates for elimination, a clause satisfied by roundness and circularity (Fig. 4). Roundness was strongly correlated with compactness ($r^2 = 0.97$), which was expected as both features are essentially a ratio of the object area to its maximum Feret diameter. Due to their equivalence, compactness was preserved. Similarly, circularity and formfactor showed a similarly high correlation ($r^2 = 0.96$) as both features are a ratio of the object's area to its perimeter, meaning either could be chosen (for this study, we chose form-

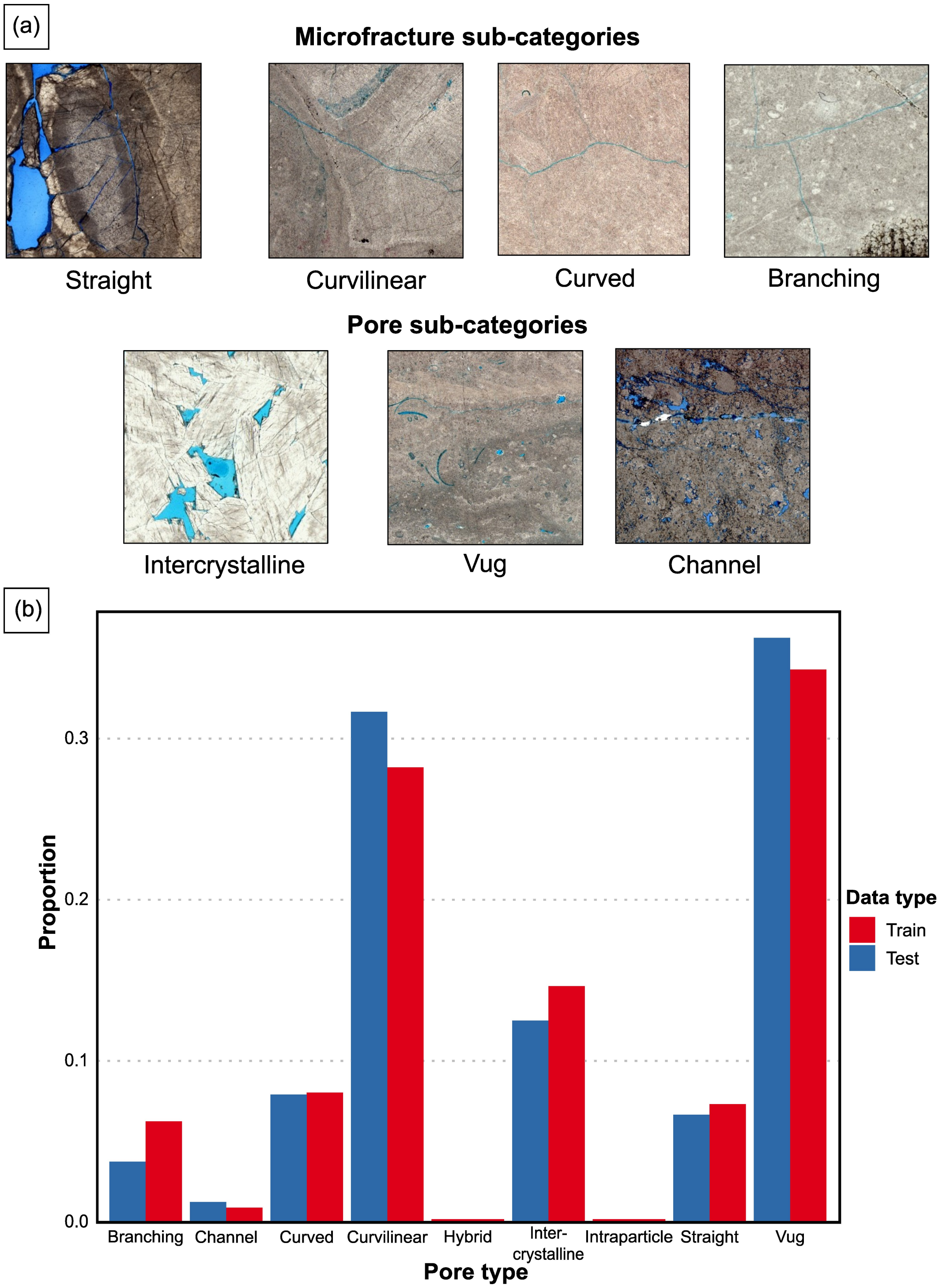

Figure 3. (a) Examples of pore and microfracture types from the dataset. (b)Proportion of pore and microfracture type in training data.

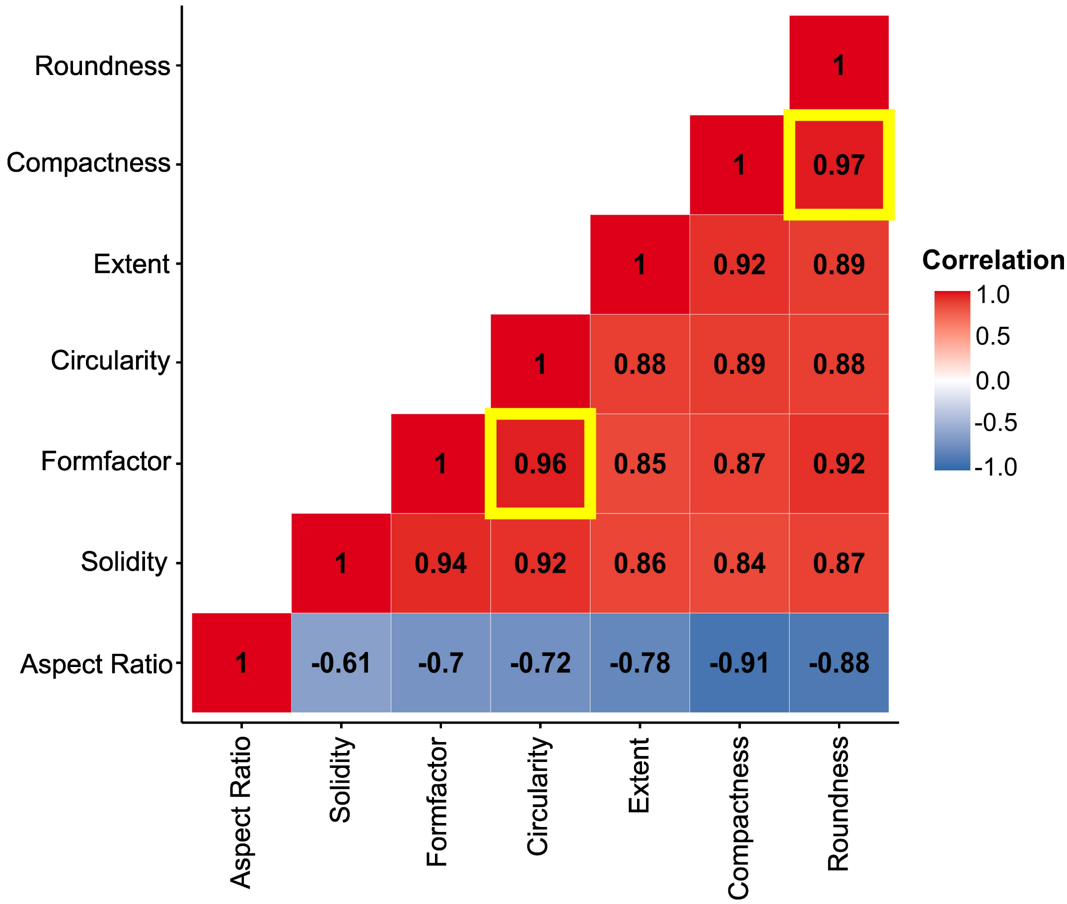

Figure 4. Correlation matrix of all input features. The yellow boxes mark the highest correlations among the features.

factor). Aspect ratio was the only exceptional feature, as it was negatively correlated with every other feature, and in particular, compactness. The final features selected were formfactor, compactness, extent, solidity, and aspect ratio. The selected features are conceptually independent of one another, with the exception of extent and solidity, which are both area-ratio variants, thus by-in-large, satisfying the independence requirement put forward by (Loncaric, 1998; Neal & Russ, 2012). A caveat to the feature selection process is that correlation analysis assumes univariate normality, an assumption that was violated by most of the features. However, since the correlation was only used to detect redundant features and was not involved in the modeling process, the impact of violating the assumptions is not an issue.

2.5.4 High Dimensional Visualization: Principal Components Analysis (PCA)

Principal Components Analysis (PCA) was used to visualize the relationships in five-dimensional space. PCA is a wholly unsupervised technique that reduces the dimensionality of data to those that explain the maximal variance (Jolliffe & Cadima, 2016; Vogelstein et al., 2021). PCA is arguably the most popular dimensionality-reduction technique (Vogelstein et al., 2021). Details on the conceptual and mathematical underpinnings of PCA can be reviewed in Jolliffe and Cadima (2016). The covariance matrix of the dataset was constructed and factorized using eigen decomposition to find its principal components. We performed PCA on the whole dataset and the labeled subset, with the same processing steps used for the training and testing data applied to both datasets.

2.5.5 Model Selection

Several supervised ML models were tested as we had no prior knowledge as to which was best suited to our problem. This practice is colloquially known as the 'No free lunch theorem' (Kuhn & Silge, 2022). The 'best' model does not necessarily mean the most accurate, but rather the model that balances accuracy with generalizability and efficiency. We tested nine models in this study: multiple logistic regression (MLR), linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), K-nearest neighbors (kNN), Naive-Bayes (NB), Random forest (RF), and three variants of Support Vector Machines (SVM); linear, radial, and polynomial. Further details on each model can be found in James et al. (2021) and Kuhn et al. (2013). These models can be broadly classified into two categories: linear and non-linear. Linear models generate linear decision boundaries in high-dimensional feature space, whereas non-linear models create non-linear decision boundaries in feature space such as polynomial, radial, or more complex non-parametric curves. All models were run using the 'caret' package in R (Kuhn, 2022).

2.5.6 Hyperparameter Optimization

Most of the tested models possessed hyperparameters that require user definition. Optimal parametrization is critical to maximize the performance of supervised models. For models without tunable hyperparameters, such as MLR, LDA, and QDA, the models were trained using 10-fold cross-validation repeated ten times with accuracy as the chosen metric. For models that contained tunable hyperparameters, a grid search technique was employed for each hyperparameter, with 10-fold cross-validation repeated ten times applied to each set of hyperparameters. The hyperparameter combination with the highest average accuracy was selected to train the final model. The list of the hyperparameters for each model (if present) and the chosen values are provided in Table S2. The hyperparameter optimization curves for each of the models are provided in the supplementary information (Fig. S6) (hyperparameter optimization was implemented using the 'trainControl' function in the 'caret' library in R).

2.5.7 Learning Curves

Learning curves were generated for the models to assess their stability and to detect any overfitting (Fig. S6). Learning curves graphically represent how well the ML model learns the classification task on incrementally larger portions of a training dataset (Kuhn et al., 2013). The typical trend is a sharp increase in training accuracy at the start as the model learns new data, eventually leading to a plateau as the model masters the task. For this study, the training and resampling increments were set at 10% of the training dataset. This meant 56 data points were used to train the model for the first run, with another 56 data points added for the second run. This incremental training was executed for ten runs till the entire training dataset was used to train the model. To check for overfitting, at each of the ten learning stages, a randomly resampled subset of the training dataset was used to test the accuracy of the model. The difference between the training and resampling curves is called the generalization gap. Typically, the lesser the gap, the more generalizable the model is considered to be (Kuhn et al., 2013).

2.5.8 Model Accuracies

As this is a binary classification study, the training and testing accuracy was measured using a confusion matrix. A confusion matrix is composed of four options: true positive (TP), false positive (FP), true negative (TN), and false negative (FN), as de-

Table 2. Confusion Matrix

True positive Positive class predicted correctly as positive	**False positive** Positive class predicted incorrectly as negative
False negative Positive class predicted incorrectly as negative	**True negative** Negative class predicted correctly as negative

fined in Table 2. Either of the classes can be designated as the positive class, with microfractures denoted as positive. Correctly predicted microfractures were classed as TP, and correctly predicted pores were classed as TN, whereas incorrect predictions for each pore type fell under FP or FN. The training and testing accuracy was calculated using (1). Whilst accuracy gives an overall picture of how accurate the model is, it does not provide information about how well the model predicted each class separately. Sensitivity, a measure of how accurately the model predicted the positive class (microfractures) (2), and specificity, a measure of how accurately the model predicted the negative class (pores), were calculated to address this deficiency (3).

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{1}$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{3}$$

2.5.9 Feature Importance

We used Shap values to evaluate the explanatory power of shape features. Initially intended to provide a means for the equitable distribution of winnings (Shapley, 1953; Lundberg & Lee, 2017), Shapley values have been appropriated from cooperative game theory into AI as a way to impute the importance of features in black-box models: a field now known as 'Explainable AI' (note that authors have coined the term 'Shap values' to differentiate from the usage of Shapley values in Game Theory: (Lundberg & Lee, 2017)). Shap values are model-agnostic and post-hoc in that they are not part of the model-building process but instead offer an external check used to explain the feature contributions to predictions. It is important to note that Shap values calculate the local importance of features, which is the importance of a particular feature to specific data points. An aggregation is performed to provide the global importance of each feature with regard to the entire dataset. For this study, both the local and global importance were measured for each model. It is also essential to acknowledge that some of the models, as an inherent aspect of their mechanics, can list the features in order of importance, namely MLR, LDA, QDA, and RF. However, we computed Shap values for all models to ensure comparison between the models.

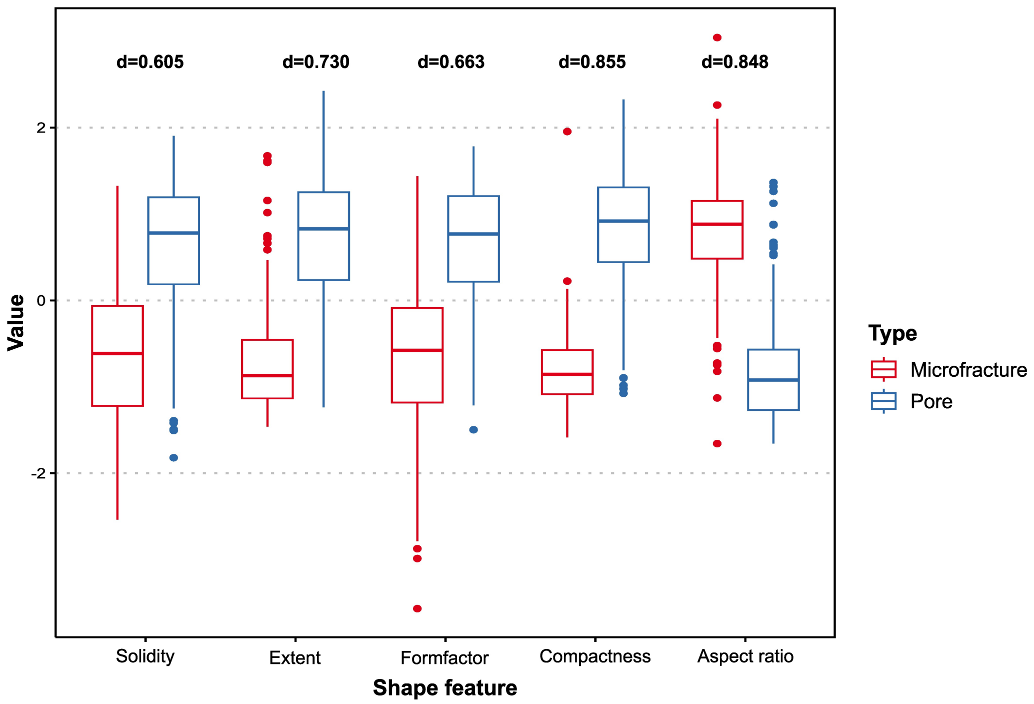

Figure 5. Differences in selected shape feature values between microfractures and pore, with the d-statistic reported for each feature. Each d-statistic was statistically significant to p¡0.001.

3 Results

3.1 Statistical Analysis of the Extracted Features

3.1.1 Univariate Distributions

The shape features for the entire dataset displayed no bimodality (Fig. S3), thus precluding any trivial assignment of decision boundaries between microfractures and pores. The lack of clear bimodality suggests the need for a high-dimensional combinatorial approach to separate the classes. However, in the labeled dataset, most of the shape features (aspect ratio, compactness, formfactor, and extent: Fig. S3) exhibited varying degrees of bimodality related to the disparate signatures of microfractures and pores (Fig. 5). However, the presence of intermediate values between the observed modes precludes the placement of straightforward decision boundaries. Visual inspection of the class populations of each shape feature suggests that compactness and aspect ratio exhibit the greatest separation between microfractures and pores, with solidity and formfactor showing the least difference, as quantified by the d-statistic from the Kolmogorov-Smirnov (K-S) test (Fig. 5).

3.1.2 PCA

The PCA biplot in the PC1-PC2 domain for the whole dataset (Fig. 6a) shows no discernable grouping but rather resembles a dense, compact cloud. The lack of separation is noteworthy, provided that PC1 and PC2 account for 93.76% of the variation in the data. The PCA visualizations containing the labeled data (Fig. 6b-c) show that the pores cluster in the direction of compactness, formfactor, and the area ratios (solidity

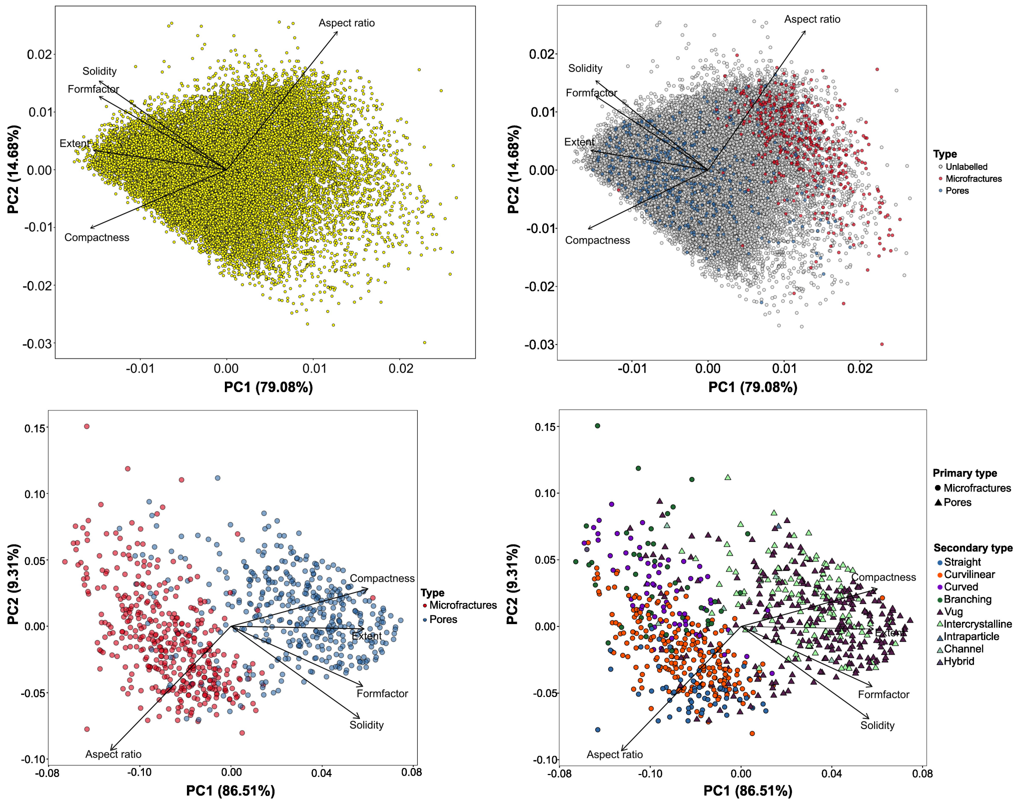

Figure 6. (a) Unlabelled PCA biplot with no separation between the datapoints. (b) PCA Biplot of the overall dataset with the labelled data overlaid. (c) PCA biplot of the labelled data. (d) Biplot of the labelled data with the secondary labels indicated.

and extent). Conversely, the more elongated microfractures cluster slightly away from the pores in the opposing direction of the aforementioned features, but in the direction of aspect ratio. It is also apparent that labeled microfractures offer a more tightly concentrated cluster, whereas the pores are more widely dispersed, with some pores overlapping within the microfractures cluster. There is also a noticeable separation between the loadings of the selected shape features, which supports the notion of independence previously alluded to. The separation of extent and solidity suggests that both features are potentially informative despite being similar area ratios.

The clustering of the labeled microfractures and pores becomes more evident when PCA is performed on the labeled dataset (Fig. 6c). PC1 and PC2 now explain a marginally higher proportion of the variance in the data (95.82%). Based upon the directions of the feature loadings, compactness and aspect ratio separate the two classes into two clusters. Furthermore, solidity and formfactor appear to extend both classes, but not sufficiently to form new clusters. This intra-class extension is further highlighted in Fig. 6d, where the datapoints are denoted by their secondary labels. In terms of pores, the two dominant pore types, intercrystalline, and vugs, show considerable overlap with no visible trend. Conversely, microfractures show a slightly discernible trend where the straight sub-class is concentrated at the base of the microfracture cluster (in the direction of increasing solidity and formfactor), and the branching and curved sub-classes concentrated near the top (in the direction of decreasing solidity and formfactor), with the curvilinear occupying the central portion of the variable space.

3.2 Supervised Machine Learning

Based on the learning curves (Fig. S6), all models show a narrow generalization gap, which indicates a lack of overfitting, except random forest, which showed overfitting to the training data (as the training accuracy was a constant 100%). In addition, most models appear to stabilize at roughly 300 data points, which points to the sufficiency of the training data for the models to learn the classification task. Another significant finding is that the linear models displayed stability and generalizability despite the lack of multi-variate normality within the training data.

Table 3. Training and testing accuracies for the supervised models

Model	Train Acc.	Train Kappa	Test Acc.	Test Kappa	95% Lower CI*	95% Upper CI*	Sens.	Spec.
MLR	94.48	88.96	90.00	80.00	85.49	93.49	96.67	83.33
LDA	94.00	88.00	89.58	79.17	85.01	93.14	97.50	81.67
QDA	94.29	88.57	90.83	81.67	86.45	94.17	97.50	84.17
kNN	94.70	89.39	90.00	80.00	85.49	93.49	94.17	85.83
NB	93.64	87.29	89.58	79.17	85.01	93.14	95.83	83.33

*CI: Confidence Interval, Sens.: Sensitivity, Spec.: Specificity

3.2.1 *Training Accuracy*

All supervised models performed highly accurately, with a strikingly narrow envelope of 93.64% to 94.63% (Table 3). To facilitate comparison between the models, the upper and lower performance bounds were measured by resampling the same training data for each model. All models perform identically, with no apparent differences between the linear and non-linear supervised models.

3.2.2 *Testing Accuracy*

The excellent performances of the models on the training data were also reflected in the testing data. Testing accuracies were only slightly lower than those of the training set and had a similarly narrow performance envelope of 89.58% to 90.83%. All models appeared to detect microfractures with greater accuracy than pores, with testing sensitivities exceeding 95%, while specificities were capped at 86%. Furthermore, the ROC curves of all the models in Fig. 7a show Area Under Curve (AUC) values > 0.95 with no observable differences between them. Despite the conceptual differences between the models, similarities in performance strongly suggest that each model's decision boundaries are similar and linear.

Despite the overall excellent performance of the models, there were systematic misclassifications. To better understand the misclassifications per model, the predicted microfracture probability of all the test data objects was derived for the sub-classes of microfractures and pores, as shown by the Polynomial SVM example in Fig. 8 (the plots for the other models are shown in Fig. S8). Most microfracture types are well above the 50% threshold across all models and, therefore, not likely to be predicted as pores. However, the branching sub-class shows the widest range of probabilities, dropping below 50% into pore prediction space in some cases. Amongst the pore types, vugs are the only class that spans nearly the entire probability range and are, therefore, responsible for the significantly lower specificities of the models. Upon closer examination, the vugs that cross the 50% threshold are dominantly bivalve molds (Fig. 8), which strongly resemble curvilinear microfractures.

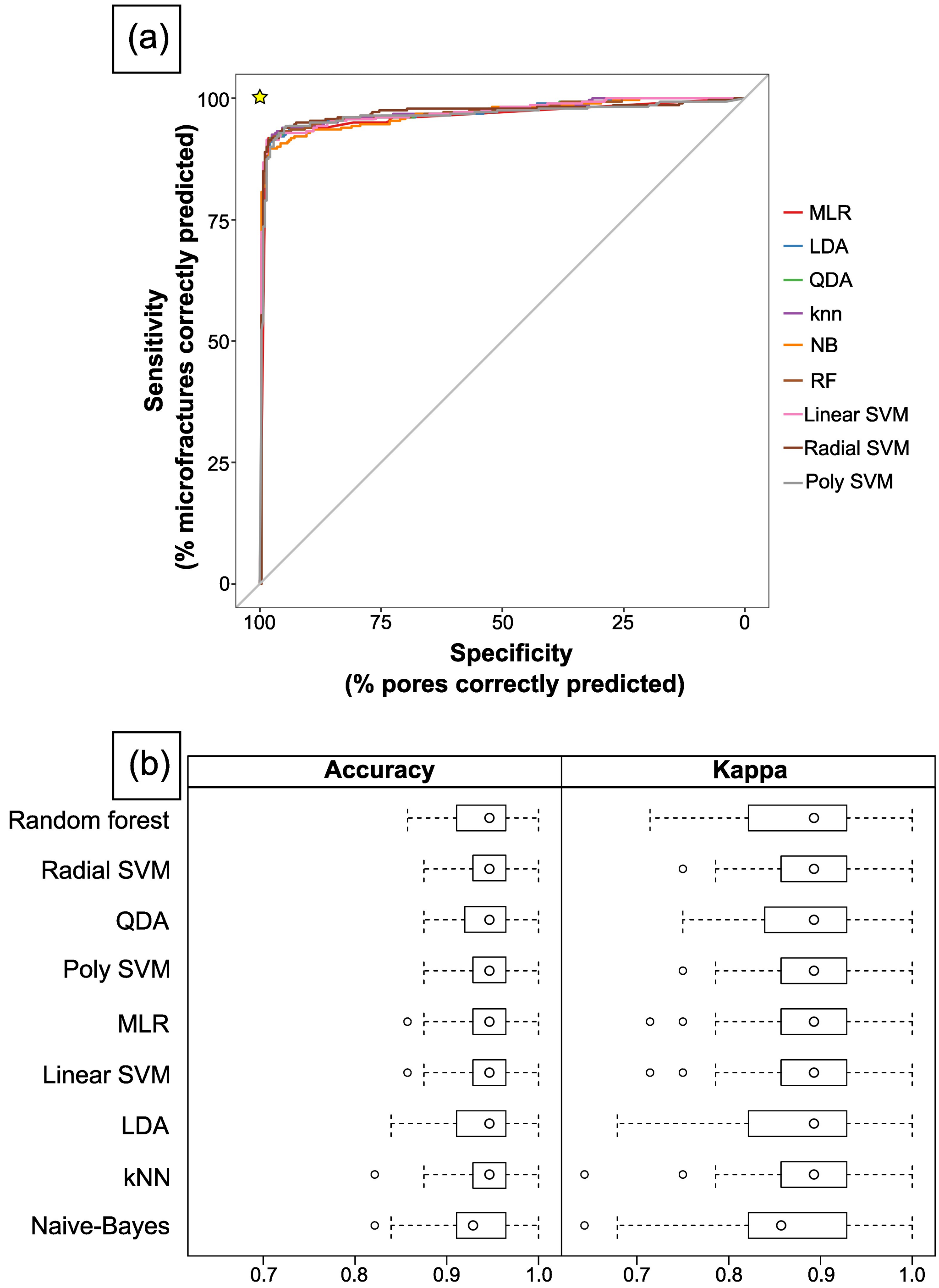

Figure 7. (a) ROC curves for all the tested models. All models show exceptionally high sensitivities and specificities across all probability thresholds. (b) Calibration curves for all the tested models. Same color scheme as (a). (c) Boxplot of training accuracies with confidence intervals derived from identical resampling.

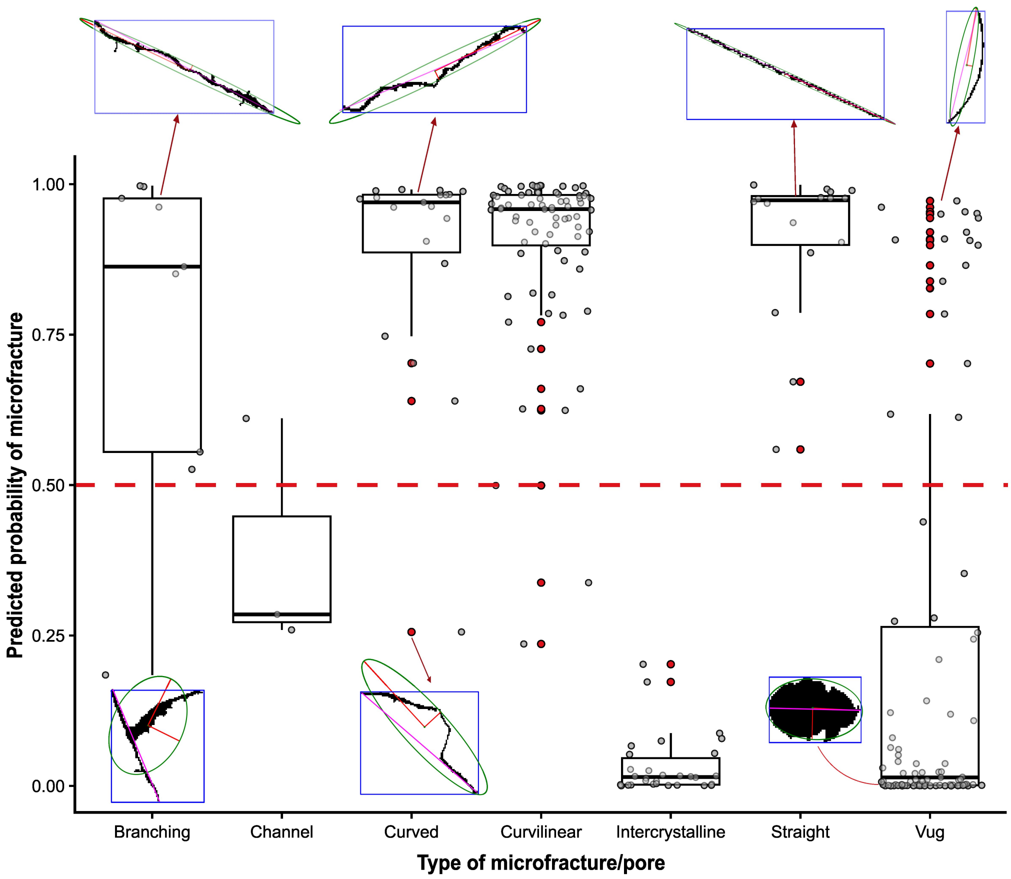

Figure 8. Microfracture prediction probability for each pore and microfracture type for Polynomial SVM model. Example masks of pore types are provided to illustrate the variation per class. The green ellipse represents the best-fitting ellipse with the red lines are the major and minor axes of the ellipse. The pink line represents the maximum Feret diameter. The blue box represents the bounding box.

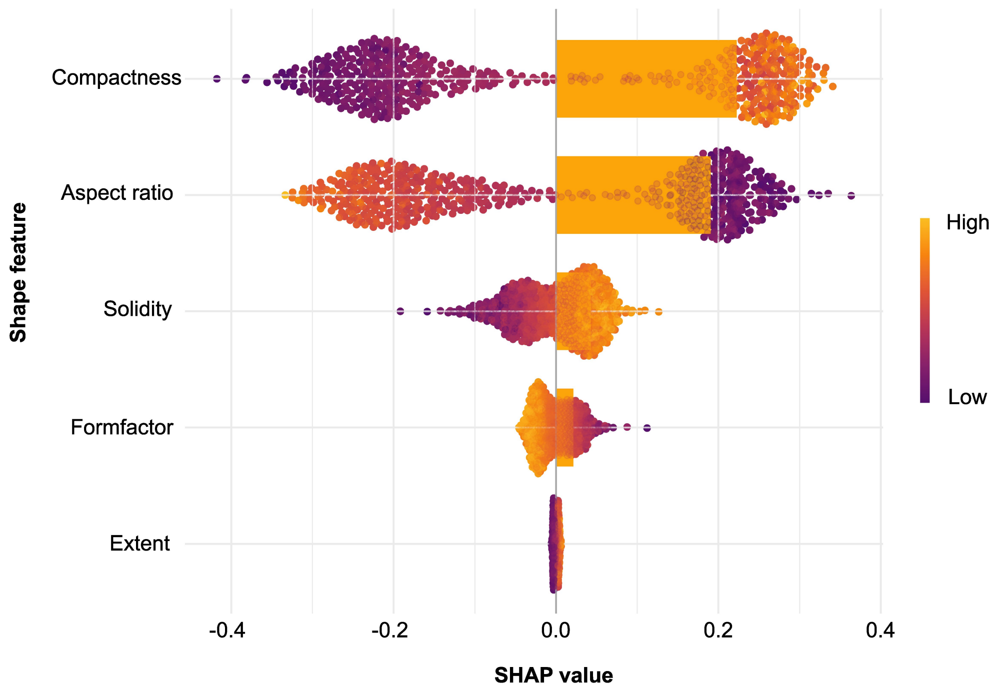

Figure 9. Shap values per feature for the Polynomial SVM model. The points represent local importance, and the bars represent global importance. The features are ordered by global importance.

3.2.3 Feature Importance

Shap plots ranking feature importance for all models are shown in Fig. S9 with only Polynomial SVM presented in Fig. 9 as a representative case. Feature rankings per model are listed in Table 4. Compactness was consistently the most important feature across models, while aspect ratio was the second-most important feature in most models tested (i.e., seven out of the nine), with MLR and LDA serving as the exceptions. Solidity was the third most important feature for most models, except for MLR and LDA (2), QDA (4), and Naive-Bayes (5). Solidity and formfactor appear to interchange positions in QDA and Naive-Bayes, which could be explained by their close correlation seen in the PCA biplot in Fig. 8b. The shape feature with the least contribution to most models is extent. It is also apparent that the models fall into three broad groups in terms of the feature importance profiles. The first group includes MLR and LDA, the second group includes the majority of the models, such as RF, KNN, linear SVM, radial SVM, and polynomial SVM, and the third group consists of QDA and Naive-Bayes.

Table 4. Rankings of the shape feature importance per model

	MLR	LDA	QDA	NB	kNN	RF	LSVM	RSVM	PSVM
Compactness	1	2	1	1	2	1	1	1	1
Aspect ratio	2	1	4	2	1	2	2	2	2
Solidity	5	3	2	5	3	3	3	3	3
Formfactor	3	4	3	4	4	4	4	4	4
Extent	4	5	5	3	5	5	5	5	5

4 Discussion

4.1 Performance of the Supervised ML models

The excellent performance of all the tested supervised ML models shows their efficacy in the presented classification task, in similitude to the high accuracies of supervised pore-type classification reported in the related literature (Table S1). However, a straight comparison with the related literature is impossible due to the difference in the predicted classes. The equivalent performance of both linear and non-linear ML models indicates ample separation between the microfractures and pores in the feature space, and the decision boundary was likely linear, thus posing a relatively simple classification task. Notably, this separation is discernable in the PCA biplot for the labeled data (Fig. 6c). Furthermore, all models contained errors related to the misclassification of bivalves as microfractures, indicating that the models did not fit complex, non-linear decision boundaries through the microfractures cluster.

4.2 The importance of compactness and aspect ratio in the labelled dataset

The importance of compactness and aspect ratio in creating discernable separation is evident from the PC1-PC2 visualization of the labeled data (Fig. 6c). Both features also ranked the highest amongst the shape features across most of the ML models based on Shap values (Fig. 9 and Fig. S9). However, compactness consistently outranked aspect ratio across most models, which is perhaps counter-intuitive given the popularity of aspect ratio as a unique identifier for microfractures in the geological community (Table S1). To better understand the ranking, the aspect ratio of an object, using the best-fitting ellipse, essentially strips the object of its natural shape by assuming that two orthogonal axes can adequately represent it. We observe in Fig. 8 that best-fitting ellipses are reasonably faithful to the geometries of the more linear microfracture types (straight and curvilinear). In contrast, more curved or branched microfractures diverge from the low aspect ratio character and start to approach more pore-like values. Fig. 6d displays this to some extent, as the curving and branching microfractures are slightly closer to the pores than the straight variety. Conversely, compactness uses the original area of the object and only approximates its maximum length (the Feret diameter), which is a reasonably robust measure of object length and approximately equivalent to the major axis of the best-fitting ellipse. In addition, compactness places less weight on the area of the object and more emphasis on its maximum length: a construct that works well in the context of microfractures as they have significantly smaller areas than most similarly sized pores and always contain an outsized axis, except for a subset of branching microfractures. We note that any feature that adequately captures the salient characteristics of microfractures, namely the elongation and relatively narrow aperture, can contribute significantly to model performance. We also note that extent proved to be the least informative across all models. The lack of information can be attributed to its sensitivity to rotation, as illustrated in Fig. 8, where the same object can have different bounding boxes based on its orientation. Therefore, extent violates the rotation-invariance requirement of shape features (Loncaric, 1998). While extent contains information on the complexity of the pore (as more complex pores only take up a smaller portion of the bounding box), the rotation sensitivity means that solidity is a better replacement information-wise.

4.3 The weaknesses of the approach when extended to the global dataset

The results of the study indicate that the classification of microfractures and pores is a simple problem, which conforms to the visual perception that these pore types are separable by simple geometric features alone (Z. Wang et al., 2022). However, whether the labeled dataset of 800 points used in this study adequately represents the global dataset of 20,060 pores is questionable. Fig. 6a-6c highlights the major differences between both sets of data, with the unlabelled data showing none of the separation seen in the labeled data, thus strongly indicating that the classification is not straightforward. The difference between the global and labeled datasets can be attributed to two main factors: geological complexity and technical considerations.

The complexity of carbonate pore types is well-known (Ehrenberg, 2022). Dissolution and cementation are spatio-temporally variable processes controlled by a myriad of depositional and diagenetic agents, which typically result in complex pore morphologies that often do not fit conveniently into classification schemes. The most popular of the pore-typing schemes, Choquette and Pray (1970), and Lucia (1983, 1995), do not contain morphology as a diagnostic attribute for this reason. To further highlight pore complexity, intercrystalline pores, and vugs overlap significantly in the PC1-PC2 space (Fig. 6d) despite their contrasting origins attributable to cementation and dissolution, respectively. In addition, microfractures can develop complex morphologies (Fig. 8) based on the heterogeneity of the rock and the stress regimes acting therein.

Further to this, the non-unique nature of simple shape features used in this study and the related literature (Abedini et al., 2018; Borazjani et al., 2016; Ghiasi-Freez et al., 2012; Mollajan et al., 2016; Z. Wang et al., 2022) could not adequately separate the pore types in hyperspace, as illustrated in Fig. 6a-d. These features can be informative for idealized objects where microfractures are mostly linear to curvilinear and pores are mostly equant. However, such scenarios are rare in the carbonate realm, and the continued reliance upon simple feature sets will likely produce dense point clouds for which classification is problematic.

4.4 The weaknesses of the approach when extended to the global dataset

4.4.1 Biased Sampling

Selection bias during the sampling phase is a likely cause for the excellent separation in the labeled data. Operator discretion was required during the random sampling procedure to filter out noise, such as microporous patches or pores below the feature resolution. While this mitigated the noise fed into the models, it also meant that the most characteristic pores would be selected, thereby compromising the objectivity of the sampling procedure. Data curation is a typical stage for pore typing studies, often resulting in overly optimistic results in supervised ML (Table S1). Comparatively, most other related studies use at most 250 data points for labeling, while we used 800.

It is evident that studies claiming excellent performance of supervised ML for pore typing have not fully considered the true complexity of the task and instead report the results of highly curated datasets (Abedini et al., 2018; Borazjani et al., 2016; Ghiasi-Freez et al., 2012; Mollajan et al., 2016; Z. Wang et al., 2022). We expect that this research avenue will continue to grow exponentially given the importance of automated pore-typing for a multitude of value-generating processes, mainly as we are well into the era of big data. Besides data curation, most related studies have only used a fraction of our ground truth size to build their models, which cannot be considered representative and will only exacerbate model accuracies (Abedini et al., 2018; Borazjani et al., 2016; Ghiasi-Freez et al., 2012; Mollajan et al., 2016; Z. Wang et al., 2022).

4.4.2 Possible weaknesses within the projection method

Another potential explanation for the lack of separation within the unlabelled data is the problematic nature of PCA with respect to the visualization of the feature space. While PCA is the most popular dimensionality-reduction approach within the scientific literature, it is also the weakest in projecting the true distances between points in 2D (Van Der Maaten et al., 2009; Thrun, 2018). In essence, large distances between points in feature space may appear close in the 2D-projected PCA space as PCA only rotates the data points to the axis containing the greatest variance. Unlike non-linear projection methods, such as Connected Components Analysis (CCA), t-distributed Stochastic Neighbor Embedding (t-SNE), and Multi-dimensional Scaling (MDS), PCA does not disaggregate the data into clusters (Van Der Maaten et al., 2009; Thrun, 2018; Thrun & Ultsch, 2021). Therefore, PCA would unlikely display clusters unless the feature space already contains appreciable clustering within the higher dimensions. Hence, it can be argued that the unlabelled feature space may contain clusters by pore type that are collapsed into one another within the PCA space. It should be noted, however, that the density-based DBSCAN method only showed one cluster for the unlabelled data (Fig. S4b), and k-means only managed to bisect the cloud through its centroid (Fig. S4a). Both results are independent of the projection and suggest that there is no discernible separation between the classes in the global feature space, which makes the use of any projection method moot for this case.

4.4.3 Dataset Size

Another factor that may have contributed to disparities in separability between the labeled and unlabeled data is the limited size of the dataset (18 images / 20060 objects), which cannot be considered representative of carbonates. Several pore types commonly observed in carbonate studies, such as interparticle pores, intraparticle molds, and channels, were limited in quantity, meaning that random sampling emphasized the more dominant intercrystalline pores and vugs. Including the former pores would potentially have resulted in a more complex feature space in the labeled dataset and be more representative of the range of pore types observed within carbonate rocks. Indeed, even the observed spectrum of pore types within the 18 thin sections studied herein was not fully representative, as only 2% of the available pores were selected as ground truth compared to approximately 90% in the case of microfractures, thereby making this study more representative of the latter. Barring a community-wide effort, scant ground truth datasets for pore typing will likely continue to be a significant bottleneck for quantitative pore typing studies in carbonate lithologies.

4.4.4 Fragmentation of microfractures

Another likely cause for the separation in the labeled data was the microfractures' fragmentation due to the scans' poor resolution. Several curved and branching microfractures were fragmented into smaller, more linear segments, resulting in a disproportionate number of linear and curvilinear microfractures (Fig. S1). This over-simplification of complex microfracture networks masked the true complexity of the feature space. The geometric complexity of microfractures would be honored more accurately with higher-resolution scans, allowing the power of supervised ML models to be benchmarked more effectively. Spatial aliasing of fractures from image datasets is a ubiquitous issue related to their characterization (Seers & Hodgetts, 2014; Biber et al., 2018). We expect that the related literature also faced similar challenges related to resolution-dependent censoring of fracture networks reported herein, though it did not address it explicitly.

4.5 Study Design Issues in Related Studies

However, the larger problem with the related studies is that they bypass the separation of microfractures and pores and directly classify pores into their sub-classes (Table S1). We show that there is heavy overlap between the pore types within the simple shape feature space, thus raising questions on the predictive accuracies of the proposed models in the literature. Again, the current dataset does not contain several pore types that share morphological similarities with microfractures, such as interparticle pores and channels, which would further convolute the feature space utilized for pore classification herein.

A related problem with most studies is that they do not explain the importance of the simple shape features in the ML models. The fact that all related studies re-use the same features without any explanation of their importance to the models only propagates poor practices in the field. For example, extent is commonly utilized within automated pore typing studies (Table S1). However, we report that extent was the least informative feature across all models (i.e., based on the Shap values: Fig 9 and S9), due to its sensitivity to rotation violating the rotation-invariance requirement of shape features (Loncaric, 1998). While extent contains information on the complexity of the pore, as more complex pores take up a smaller portion of the bounding box, the rotation sensitivity means that solidity offers a more attractive alternative.

Finally, most related studies lack robust supervised ML methodologies (Abedini et al., 2018; Borazjani et al., 2016; Ghiasi-Freez et al., 2012; Mollajan et al., 2016; Sharifi, 2022; Z. Wang et al., 2022). Feature selection appears to be related more to the ease of acquisition rather than any proven utility. Most studies do not undertake visualization of the data in hyperspace using PCA (Abedini et al., 2018; Borazjani et al., 2016; Ghiasi-Freez et al., 2012; Mollajan et al., 2016; Z. Wang et al., 2022), thereby obfuscating the underpinning drivers of their reported excellent model accuracies. Almost all related studies do not furnish details on hyperparameter tuning, perhaps as the default parameters produce excellent results (Abedini et al., 2018; Borazjani et al., 2016; Ghiasi-Freez et al., 2012; Mollajan et al., 2016; Z. Wang et al., 2022). Also, there needs to be more comparison across several different models, particularly with simpler classifier paradigms, to provide a baseline performance (Table S1).

4.6 Moving Forward

The classification of microfractures and pores is still a complex problem that requires attention. Given that these features are ostensibly geometric end members, it is more prudent to approach this problem prior to drawing finer distinctions in pore types using multiclass ML frameworks. Macrofracture segmentation studies follow this template with emphasis on extracting the macrofractures in microCT models by all possible means, with the other class inherently being pores (Lee et al., 2021). Ideally, enhancing the separation of microfractures and pores into natural clusters in the feature space should be prioritized. The presence of natural clusters would enable the use of unsupervised clustering models directly on the dataset or even on the dimensionally reduced projection (referred to as projection-based clustering) (Van Der Maaten et al., 2009; Thrun, 2018; Thrun & Ultsch, 2021). An unsupervised approach is scalable and has the added benefit of not requiring labeled data. However, natural clustering in the feature space is not likely using simple shape features. We hypothesize that more complex shape features such as the contour-based Fourier descriptors and region-based invariant moments (invariant Hu moments and Zernike moments) might create better separations in hyperspace, albeit with an attendant decrease in explainability of the features (Neal & Russ, 2012; Singh et al., 2021). It is also possible that in concert with more complex features, more powerful methods of dimensionality-reduction, such as CCA, MDS, and t-SNE, may enhance the presence of natural clusters for projection-based clustering (Thrun, 2018; Thrun & Ultsch, 2021). We note that a DL approach would likely offer the best results; however, to be feasible, it would require data sharing and ground truth labeling on a hitherto unprecedented scale within the geoscience community. It is pertinent to not only have a global representation of pore and microfracture types but also of a range of instruments with different acquisition parameters to ensure the generalizability of the classifiers. It would also require a community effort to find the best shape features and AI models, potentially borrowing from equivalent studies within the fields of computer vision and bioinformatics, for example, where similar applications of supervised and unsupervised machine learning towards object clustering and classification from image data is already mature (Butler et al., 2018; Chen et al., 2019; Doerr & Florence, 2020; Stafford et al., 2020; Urbanowicz et al., 2020; A. Y.-T. Wang et al., 2020). Studies utilizing limited data, such as the present study, are likely to succumb to the problems of lack of representation, selection bias, and technical issues related to the imaging process, which can be conveniently masked by overly optimistic results that cannot be translated to other datasets (Sun et al., 2009).

The findings of this study serve as a benchmark for ideal datasets with limited scope of pore types. Even simple linear models such as MLR and LDA can perform excellently within such scenarios. However, we argue that the overly optimistic results from related supervised ML studies using only simple shape features are more reflective of the sampling process than the underlying geometric complexity of the pore system. We also emphasize the methodological requirement of measuring the feature importance based on the PCA loadings and their Shap values per model. This essential exploratory data analysis step will ensure that only the most important features will be carried forward into future studies rather than needlessly recycled.

5 Conclusions

All the tested supervised models performed excellently in discriminating between microfractures and pores, with testing accuracies approaching 90% for all models. Notably, all tested supervised models exhibited near identical performance, indicating a significant separation between the two classes in hyperspace such that a linear boundary was adequate. The presence of a linear decision boundary was further supported by PCA visualization of the hyperspace and the systematic misclassification of bivalve molds as microfractures. However, upon comparing the feature spaces of the labeled data and the overall dataset, it is apparent that the labeled feature space presented a highly sanitized version of the larger dataset despite efforts toward the development of an objective sampling scheme. The sanitized dataset converted a complex problem requiring complex nonlinear decision boundaries to a simple, linearly separable problem. While our study can provide a useful benchmark for those that contain more idealized datasets with limited microfracture and pore types, we demonstrate that the pore-typing problem is more complex than postulated by the related literature. Finally, we report that, contrary to expectations, compactness contributed more towards the ML classification of microfractures from pores than aspect ratio, as compactness only approximates one measure of the object compared to the two metrics approximated by aspect ratio. These results serve as a useful template for future studies on this first-order challenge of separating microfractures and pores and on higher-order challenges involving more complex multiclass pore typing.

6 Open Research

The image data used for the classification in the study and the R code developed are published at the GitHub repository for this study via https://github.com/issacsujay92/Microfractures-And-Pores-ML with no restriction on usage. The entire code was developed in R (version 4.2.1) (R Core Team, 2022) using the RStudio IDE. Figures were made using ggplot2 package (Wickham, 2016). The ML models were run using 'caret' version 6.0.93 (Kuhn, 2022). Data analytics and visualizations were implemented using the following packages: 'tidyverse' (Wickham et al., 2019), 'MASS' (Venables & Ripley, 2002), 'factoextra' (Kassambara & Mundt, 2020), 'FactoMineR' (Lê et al., 2008), 'ggfortify' (Tang et al., 2016), 'GGally' (Schloerke et al., 2021), 'klaR' (Weihs et al., 2005), and 'reshape2' (Wickham, 2007). Model performance evaluation was implemented using the 'MLeval' package (John, 2020). fastshap (Greenwell, 2021), and shapviz (Mayer, 2023) were essential to implementing and visualizing the Shap values for the ML models.

Acknowledgments

The authors thank Dr. Siddarth Misra and Dr. Jim Ji for guiding the project from the inception. Dr. David Bapst and Dr. Christina Belanger are thanked for discussions during early versions of the draft and for their excellent course on implementing R for geological datasets. Issac would also like to thank Dr. Ankita Singh for enlightening discussions on the technical aspects of the study and the bigger picture implications. The financial assistance of Qatar Foundation, the Qatar National Research Fund (NPRP12S-0302-190194), and Total Energies Qatar are gratefully acknowledged by the authors.

References

Abedini, M., Ziaii, M., Negahdarzadeh, Y., & Ghiasi-Freez, J. (2018). Porosity classification from thin sections using image analysis and neural networks including shallow and deep learning in jahrum formation. *Journal of Mining and Environment*, *9*(2), 513–525.

Ansari, M. Y., Abdalla, A., Ansari, M. Y., Ansari, M. I., Malluhi, B., Mohanty, S., ... others (2022). Practical utility of liver segmentation methods in clinical surgeries and interventions. *BMC medical imaging*, *22*(1), 1–17.

Ansari, M. Y., Yang, Y., Balakrishnan, S., Abinahed, J., Al-Ansari, A., Warfa, M., ... others (2022). A lightweight neural network with multiscale feature enhancement for liver ct segmentation. *Scientific reports*, *12*(1), 14153.

Artrith, N., Butler, K. T., Coudert, F.-X., Han, S., Isayev, O., Jain, A., & Walsh, A. (2021). Best practices in machine learning for chemistry. *Nature chemistry*, *13*(6), 505–508.

Biber, K., Khan, S. D., Seers, T. D., Sarmiento, S., & Lakshmikantha, M. (2018). Quantitative characterization of a naturally fractured reservoir analog using a hybrid lidar-gigapixel imaging approach. *Geosphere*, *14*(2), 710–730.

Bishop, C. M. (2006). *Pattern recognition and machine learning* (Vol. 4) (No. 4). Springer.

Blaschke, T. (2010). Object based image analysis for remote sensing. *ISPRS journal of photogrammetry and remote sensing*, *65*(1), 2–16.

Borazjani, O., Ghiasi-Freez, J., & Hatampour, A. (2016). Two intelligent pattern recognition models for automatic identification of textural and pore space characteristics of the carbonate reservoir rocks using thin section images. *Journal of Natural Gas Science and Engineering*, *35*, 944-955. Retrieved from `https://www.sciencedirect.com/science/article/pii/S1875510016306886` doi: https://doi.org/10.1016/j.jngse.2016.09.048

Buades, A., Coll, B., & Morel, J.-M. (2011, September). Non-local means denoising. *Image Processing On Line*, *1*, 208–212. Retrieved from `https://doi.org/10.5201/ipol.2011.bcm_nlm` doi: 10.5201/ipol.2011.bcm_nlm

Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O., & Walsh, A. (2018). Machine learning for molecular and materials science. *Nature*, *559*(7715), 547–555.

Chawla, N. V., Japkowicz, N., & Kotcz, A. (2004, jun). Editorial: Special issue on learning from imbalanced data sets. *SIGKDD Explor. Newsl.*, *6*(1), 1–6. Retrieved from `https://doi.org/10.1145/1007730.1007733` doi: 10.1145/1007730.1007733

Chen, P.-H. C., Liu, Y., & Peng, L. (2019). How to develop machine learning models for healthcare. *Nature materials*, *18*(5), 410–414.

Choquette, P. W., & Pray, L. C. (1970, 02). Geologic Nomenclature and Classification of Porosity in Sedimentary Carbonates1. *AAPG Bulletin*, *54*(2), 207-250. Retrieved from `https://doi.org/10.1306/5D25C98B-16C1-11D7-8645000102C1865D` doi: 10.1306/5D25C98B-16C1-11D7-8645000102C1865D

Darbon, J., Cunha, A., Chan, T. F., Osher, S., & Jensen, G. J. (2008, May). Fast nonlocal filtering applied to electron cryomicroscopy. In *2008 5th IEEE inter-*

national symposium on biomedical imaging: From nano to macro. IEEE. Retrieved from https://doi.org/10.1109/isbi.2008.4541250 doi: 10.1109/isbi.2008.4541250

Doerr, F. J., & Florence, A. J. (2020). A micro-xrt image analysis and machine learning methodology for the characterisation of multi-particulate capsule formulations. *International journal of pharmaceutics: X*, *2*, 100041.

Ehrenberg, S. (2022). The etiology of carbonate porosity. *AAPG Bulletin*, *106*(12), 2351–2386.

Galar, M., Fernández, A., Barrenechea, E., Bustince, H., & Herrera, F. (2011). An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes. *Pattern Recognition*, *44*(8), 1761-1776. Retrieved from https://www.sciencedirect.com/science/article/pii/S0031320311000458 doi: https://doi.org/10.1016/j.patcog.2011.01.017

Ghiasi-Freez, J., Soleimanpour, I., Kadkhodaie-Ilkhchi, A., Ziaii, M., Sedighi, M., & Hatampour, A. (2012). Semi-automated porosity identification from thin section images using image analysis and intelligent discriminant classifiers. *Computers & geosciences*, *45*, 36–45.

Graf, R., Zeldovich, M., & Friedrich, S. (2022). Comparing linear discriminant analysis and supervised learning algorithms for binary classification—a method comparison study. *Biometrical Journal*.

Greener, J. G., Kandathil, S. M., Moffat, L., & Jones, D. T. (2022). A guide to machine learning for biologists. *Nature Reviews Molecular Cell Biology*, *23*(1), 40–55.

Greenwell, B. (2021). fastshap: Fast approximate shapley values [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=fastshap (R package version 0.0.7)

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263-1284. doi: 10.1109/TKDE.2008.239

Hemes, S., Desbois, G., Urai, J. L., Schröppel, B., & Schwarz, J.-O. (2015). Multi-scale characterization of porosity in boom clay (hades-level, mol, belgium) using a combination of x-ray μ-ct, 2d bib-sem and fib-sem tomography. *Microporous and mesoporous materials*, *208*, 1–20.

James, G., Witten, D., Hastie, T., Tibshirani, R., James, G., Witten, D., … Tibshirani, R. (2021). Statistical learning. *An introduction to statistical learning: with applications in R*, 15–57.

John, C. R. (2020). Mleval: Machine learning model evaluation [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=MLeval (R package version 0.3)

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences*, *374*(2065), 20150202.

Kassambara, A., & Mundt, F. (2020). factoextra: Extract and visualize the results of multivariate data analyses [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=factoextra (R package version 1.0.7)

Kuhn, M. (2022). caret: Classification and regression training [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=caret (R package version 6.0-93)

Kuhn, M., & Johnson, K. (2019). *Feature engineering and selection: A practical approach for predictive models.* Chapman and Hall/CRC.

Kuhn, M., Johnson, K., et al. (2013). *Applied predictive modeling* (Vol. 26). Springer.

Kuhn, M., & Silge, J. (2022). *Tidy modeling with r.* " O'Reilly Media, Inc.".

Lê, S., Josse, J., & Husson, F. (2008). FactoMineR: A package for multivariate analysis. *Journal of Statistical Software*, *25*(1), 1–18. doi: 10.18637/jss.v025.i01

Lee, D., Karadimitriou, N., Ruf, M., & Steeb, H. (2021). Detecting micro fractures with x-ray computed tomography. *arXiv preprint arXiv:2103.12821*.

Legland, D., Arganda-Carreras, I., & Andrey, P. (2016). Morpholibj: integrated library and plugins for mathematical morphology with imagej. *Bioinformatics*, *32*(22), 3532–3534.

Li, B., Tan, X., Wang, F., Lian, P., Gao, W., & Li, Y. (2017). Fracture and vug characterization and carbonate rock type automatic classification using x-ray ct images. *Journal of Petroleum Science and Engineering*, *153*, 88–96.

Loncaric, S. (1998). A survey of shape analysis techniques. *Pattern recognition*, *31*(8), 983–1001.

Lucia, F. J. (1983, 03). Petrophysical Parameters Estimated From Visual Descriptions of Carbonate Rocks: A Field Classification of Carbonate Pore Space. *Journal of Petroleum Technology*, *35*(03), 629-637. Retrieved from `https://doi.org/10.2118/10073-PA` doi: 10.2118/10073-PA

Lucia, F. J. (1995, 09). Rock-Fabric/Petrophysical Classification of Carbonate Pore Space for Reservoir Characterization1. *AAPG Bulletin*, *79*(9), 1275-1300. Retrieved from `https://doi.org/10.1306/7834D4A4-1721-11D7-8645000102C1865D` doi: 10.1306/7834D4A4-1721-11D7-8645000102C1865D

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.

Lønøy, A. (2006, 09). Making sense of carbonate pore systems. *AAPG Bulletin*, *90*(9), 1381-1405. Retrieved from `https://doi.org/10.1306/03130605104` doi: 10.1306/03130605104

Mayer, M. (2023). shapviz: Shap visualizations [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=shapviz` (R package version 0.5.0)

McCreesh, C. A., Ehrlich, R., & Crabtree, S. J. (1991). Petrography and reservoir physics ii: relating thin section porosity to capillary pressure, the association between pore types and throat size. *AAPG bulletin*, *75*(10), 1563–1578.

Mollajan, A., Ghiasi-Freez, J., & Memarian, H. (2016). Improving pore type identification from thin section images using an integrated fuzzy fusion of multiple classifiers. *Journal of Natural Gas Science and Engineering*, *31*, 396–404.

Neal, F. B., & Russ, J. C. (2012). *Measuring shape*. CRC Press.

R Core Team. (2022). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Rabbani, A., Fernando, A., Shams, R., Singh, A., Mostaghimi, P., & Babaei, M. (2021). Review of data science trends and issues in porous media research with a focus on image-based techniques. *Water Resources Research*, *57*(10), e2020WR029472.

Schloerke, B., Cook, D., Larmarange, J., Briatte, F., Marbach, M., Thoen, E., ... Crowley, J. (2021). Ggally: Extension to 'ggplot2' [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=GGally` (R package version 2.1.2)

Seers, T. D., & Hodgetts, D. (2014). Comparison of digital outcrop and conventional data collection approaches for the characterization of naturally fractured reservoir analogues. *Geological Society, London, Special Publications*, *374*(1), 51–77.

Seers, T. D., & Hodgetts, D. (2016). Extraction of three-dimensional fracture trace maps from calibrated image sequences. *Geosphere*, *12*(4), 1323–1340.

Shapley, L. S. (1953). Stochastic games. *Proceedings of the national academy of sciences*, *39*(10), 1095–1100.

Sharifi, J. (2022). Multi-pore rock physics model: An intelligent approach for carbonate rocks. *Journal of Petroleum Science and Engineering*, *218*, 111002. Retrieved from https://www.sciencedirect.com/science/article/pii/S092041052200852X doi: https://doi.org/10.1016/j.petrol.2022.111002

Singh, A., Rabbani, A., Regenauer-Lieb, K., Armstrong, R. T., & Mostaghimi, P. (2021). Computer vision and unsupervised machine learning for pore-scale structural analysis of fractured porous media. *Advances in Water Resources*, *147*, 103801.

Skalinski, M., & Kenter, J. A. M. (2015). Carbonate petrophysical rock typing: integrating geological attributes and petrophysical properties while linking with dynamic behaviour. *Geological Society, London, Special Publications*, *406*(1), 229-259. Retrieved from https://www.lyellcollection.org/doi/abs/10.1144/SP406.6 doi: 10.1144/SP406.6

Stafford, I., Kellermann, M., Mossotto, E., Beattie, R. M., MacArthur, B. D., & Ennis, S. (2020). A systematic review of the applications of artificial intelligence and machine learning in autoimmune diseases. *NPJ digital medicine*, *3*(1), 30.

Sun, Y., Wong, A. K., & Kamel, M. S. (2009). Classification of imbalanced data: A review. *International journal of pattern recognition and artificial intelligence*, *23*(04), 687–719.

Tang, Y., Horikoshi, M., & Li, W. (2016). ggfortify: unified interface to visualize statistical results of popular r packages. *R J.*, *8*(2), 474.

Thrun, M. C. (2018). *Projection-based clustering through self-organization and swarm intelligence: combining cluster analysis with the visualization of high-dimensional data*. Springer.

Thrun, M. C., & Ultsch, A. (2021). Using projection-based clustering to find distance-and density-based clusters in high-dimensional data. *Journal of Classification*, *38*, 280–312.

Urbanowicz, R. J., Suri, P., Cui, Y., Moore, J. H., Ruth, K., Stolzenberg-Solomon, R., & Lynch, S. M. (2020). A rigorous machine learning analysis pipeline for biomedical binary classification: application in pancreatic cancer nested case-control studies with implications for bias assessments. *arXiv preprint arXiv:2008.12829*.

Van Der Maaten, L., Postma, E. O., van den Herik, H. J., et al. (2009). Dimensionality reduction: A comparative review. *Journal of Machine Learning Research*, *10*(66-71), 13.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with s* (Fourth ed.). New York: Springer. Retrieved from https://www.stats.ox.ac.uk/pub/MASS4/ (ISBN 0-387-95457-0)

Vogelstein, J. T., Bridgeford, E. W., Tang, M., Zheng, D., Douville, C., Burns, R., & Maggioni, M. (2021). Supervised dimensionality reduction for big data. *Nature communications*, *12*(1), 2872.

Wang, A. Y.-T., Murdock, R. J., Kauwe, S. K., Oliynyk, A. O., Gurlo, A., Brgoch, J., … Sparks, T. D. (2020). Machine learning for materials scientists: An introductory guide toward best practices. *Chemistry of Materials*, *32*(12), 4954-4965.

Wang, Z., Ge, H., Zhou, W., Wei, Y., Wang, B., Liu, S., … Du, S. (2022). Characterization of pores and microfractures in tight conglomerate reservoirs. *International Journal of Hydrogen Energy*, *47*(63), 26901–26914.

Weihs, C., Ligges, U., Luebke, K., & Raabe, N. (2005). klar analyzing german business cycles. In D. Baier, R. Decker, & L. Schmidt-Thieme (Eds.), *Data analysis and decision support* (p. 335-343). Berlin: Springer-Verlag.

Wickham, H. (2007). Reshaping data with the reshape package. *Journal of Statistical Software*, *21*(12), 1–20. Retrieved from http://www.jstatsoft.org/v21/i12/

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from https://ggplot2.tidyverse.org

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., . . . Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. doi: 10.21105/joss.01686

JGR: Solid Earth

Supporting Information for

An object-based approach to differentiate pores and microfractures in petrographic analysis using explainable, supervised machine learning

Issac Sujay Anand John Jayachandran[1,2], Juan Carlos Laya[1], Holly Catherine Gibbs[3,4], Yemna Qaiser[2], Talha Khan[2], Mohammed Ishaq Mohammed Shoeb Ansari[5], Mohammed Yaqoob Ansari[5], Mohammed Malyah[2], Nayef Alyafei[2], Thomas Daniel Seers[2]

[1] Department of Geology & Geophysics, Texas A&M University, College Station, TX 77843, USA

[2] Department of Petroleum Engineering, Texas A&M University Qatar, Education City, Doha, Qatar

[3] Department of Biomedical Engineering, Texas A&M University, College Station, TX 77843, USA

[4] Microscopy and Imaging Center, Texas A&M University, College Station, TX 77843, USA

[5] Department of Electrical & Computer Engineering, Texas A&M University Qatar, Education City, Doha, Qatar

Contents of this file

Introduction

We expand on the methodological decisions made for our study. We also provide brief descriptions on the conceptual underpinnings of the supervised models used for our study. In addition, we include the results of the unsupervised clustering algorithms we tested (K-means and DBSCAN) as our study was limited to supervised models in scope. Univariate analytics on the simple shape features per secondary pore types is provided as well.

Text S1.

Pre-processing of the Images

Denoising

The kernel size of the non-local means filter was automatically estimated using the approach of Immerkaer (1996), with the smoothness factor maintained at a value of one for all images to limit over-smoothing of edges.

Sharpening

A standard unsharp mask radius of 1 with a mask value of 0.7 was applied for all images, using the built-in unsharp mask filter within Fiji. The pre-processed images after both denoising and sharpening are included in the dataset attached to this study.

Segmentation

To increase the connectivity of the microfractures, a more aggressive form of segmentation was pursued. The segmentation protocol implemented was performed in two phases. The first phase was manual thresholding of the blue epoxy impregnated pixels in the HSB (Hue, Saturation, Brightness) color space. The blue hue corresponding to the epoxy was delineated within 120 to 180, with the saturation unaltered. The lower threshold for brightness was decreased to accommodate all the blue epoxy, with values ranging from 30 to 255. The quality of the segmentation was evaluated visually in real time and parameters tuned accordingly. The second phase involved thresholding in the CIELAB color space. CIELAB color space is a device-independent method to objectively classify colors, where L stands for lightness, A for the continuum from red to green, and B for the continuum from blue to yellow (Mlynarczuk et al., 2013). Since the microfractures were filled with blue epoxy, the B channel was especially sensitive. The image was first converted from RGB to LAB color space using a Fiji built-in tool. The B channel was extracted and a simple contrast enhancement was needed to binarize the image. Both segmentation steps were performed independently. The post-processing pipeline was conducted on both the HSB and LAB binary masks in parallel, as shown in Fig. 2a.

Combining the Processed HSB and CIELAB Binary Masks

The post-processed binary masks from both color spaces were then added together using the Image Calculator tool in Fiji. The combination of both segmentations did not offer a significant boost in terms of pore connectivity as in both cases the segmentation results were similar, with the HSB binary mask offering visibly better results. Instead, the greatest effect was observed in microfractures as several individual microfractures which were disjointed from HSB thresholding displayed improved continuity in the composite image (Fig. S2). The microporous matrix zones and microporous grains were segmented as macropores as a byproduct of the aggressive segmentation. Additionally, the large quantities of these zones rendered manual masking impracticable. For this study, they were approximated as pores, which is not entirely unreasonable for grain molds and small patches of blue haze in terms of shape. However, larger microporous patches are among the major artifacts present in the data. Moreover, the thin sections used herein were not purposed for digital image analysis and as such contain damage of different forms such as pen markings and microsampling scratches amongst others. However, due to their limited quantities these scene artifacts were removed using manual masking.

Feature Extraction

Labelling of the binary masks was performed in Python using the 'Connected components' function with 8-connect from the OpenCV library. The 'regionprops' function from the sci-kit image library was used to extract size and shape features of each object (Table 1). Two associated features metrics that require special mention are the major and minor axes of the best-fitting ellipse. These axes were fit using the normalized second central moments of the object, which is a region-based approach. Region-based approaches are generally more robust than contour-based approaches as the area of the object is less sensitive to noise (Mulchrone and Choudhury, 2004; Neal and Russ, 2012). The 'regionprops' features are mostly related to object size. This required supplementing with shape features engineered from these size metrics. Feature engineering was performed in the R programming language based on derivations laid out in Neal and Russ (2012) and Weger (2006). Engineered features were selected based on their popularity in the geological community (Anselmetti et al., 1998; Weger, 2006; Weger et al., 2009; Norbisrath et al., 2015; Abedini et al., 2018; Borazjani et al., 2016; Ghiasi-Freez et al., 2012; Mollajan et al., 2016; Sharifi et al., 2022; Wang et al., 2022).

Clustering Algorithms

The objective of clustering algorithms is to group similar datapoints into discrete clusters. Two independent clustering algorithms: k-means and DBSCAN, were utilized to check for the presence of natural clusters in the feature space, ideally corresponding to microfractures and pores. The clustering algorithms were applied directly on the data. Hierarchical clustering was ignored for this study on conceptual and practical grounds. Conceptually, the objects do not necessarily follow a hierarchy, so this form of clustering is not appropriate. From a practical perspective, hierarchical clustering is also computationally expensive for large datasets such as the one in this study.

K-means Clustering

K-means was chosen as it is one of the most widely used clustering algorithms (James et al., 2021). As K-means is distance-based, it uses distance mapping to measure the distances between each of the 'n' datapoints to each other within the feature space. It then attempts to minimize the total inter-cluster distance of all clusters (James et al., 2021). The number of clusters K = 2 was selected since this study is a binary classification problem. K-Means clustering was implemented using the 'kmeans' function from the 'stats' library with euclidean distance mapping. The resultant clustering was visualized via PCA, as the feature space exceeded three dimensions, with the boundary of the clusters defined via their convex hull (Fig. S4a). It can be observed in Fig. S4a that the k-means algorithm essentially bisected the point cloud, which typically indicates a lack of natural clusters (Thrun, 2018; Thrun, 2021).

DBSCAN Clustering

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) was first proposed by Ester et al. (1996). The most significant advantages of this method compared to K-means are that it is not necessary to predefine the number of clusters and is significantly more robust to the presence of outliers (Schubert et al., 2017). DBSCAN attempts to classify the densest clusters with lower density collections of points potentially classed as outliers. This method contains two definable parameters: the cluster radius (epsilon), and the minimum number of points each cluster should contain to be considered a viable group (cluster density). Any point with the number of neighbors greater than the minimum points is considered a core point. Any point that does not have the threshold minimum points but is part of another core point neighborhood is designated a border point. If a point is neither core nor border, it is considered an outlier. For this study, the epsilon value ($\varepsilon = 0.4$) was determined by identifying the elbow of a 5-NN plot. The DBSCAN results in Fig. S4b show a single dense cluster with a few outliers, which supports the lack of natural clusters within the feature space.

Supervised ML

This section furnishes details on the conceptual underpinnings of each of the supervised ML models used in this study. Further details on each model can be found in Kuhn (2013) and James et al. (2021). The tested models can be broadly classified into two categories: linear and non-linear. Linear models generate linear decision boundaries in high-dimensional feature space, whereas non-linear models create non-linear decision boundaries, such as polynomial, radial, and more complex non-parametric curves.

Linear Models

Multiple Logistic regression (MLR) is designed for binary classification problems (Kuhn, 2013; James et al., 2021), where 'multiple' refers to the features used to train the model. Is based upon the concept of the logistic function, where the probability of classifying datapoints into the two classes resembles an S-shaped curve from 0 to 1. The logistic function is fit using maximum likelihood estimation. A major benefit of MLR is that it does not contain any tuning parameters, as the maximum likelihood estimate of the logistic function will provide the best possible model.

Linear Discriminant Analysis (LDA) and MLR only differ in their fitting procedure: whilst MLR uses maximum likelihood estimation for finding the best fitting model, LDA utilizes the Bayes' theorem. LDA assumes that the datapoints of each class belong to a Gaussian distribution, with all classes sharing a common covariance matrix. It is important to note that MLR does not require that the datapoints be drawn from multivariate Gaussian distributions and can potentially outperform LDA if the assumptions are unmet (James et al., 2021).

Non-Linear Models

Quadratic Discriminant Analysis (QDA) is similar to LDA in that it assumes the datapoints have been drawn from multivariate Gaussian distributions with the exception that each of the classes is considered to have its own covariance matrix. This difference results in a quadratic decision boundary. The greater flexibility in shape means that QDA has lower bias compared to LDA, although this typically comes at a cost of higher variance (James et al., 2021). Like MLR and LDA, QDA does not possess any tuning parameters.

Naive-Bayes, just like LDA, and QDA, is part of a family of models based on the Bayes' theorem. The 'Naive' refers to the classifier's assumption that each of the input features are uncorrelated to each other, which in this study and most other cases, is a flawed assumption. The Naive-Bayes implementation used in this study has three tunable parameters: namely the Laplace correction, distribution type, and the bandwidth adjustment.

K-nearest neighbors (KNN) is one of the simplest models commonly deployed for classification (Murphy, 2022). KNN classifies a datapoint as belonging to a certain class based on the classes of the datapoints closest to it. Thus, it does not depend on the underlying distributions of the classes, and is therefore non-parametric (James et al., 2021). The only tuning parameter for KNN is the number of neighbours (K) to each datapoint. The number of neighbors has to be odd to ensure a tiebreaker in the case of binary classification. Choosing the optimum K is non-trivial. If the number of neighbors chosen is too low, then there is a greater chance of. Conversely, if the number of neighbors is too high then decision boundaries are too general, potentially leading to underfitting.

Random Forest (RF) is an ensemble method that is based on aggregating the votes of several decision trees (Breiman, 2001; Kuhn, 2013). For each split of a decision tree, RF only allows a subset of the features to be selected. This restriction ensures that features that strongly influence the datapoints will not be preferred as several trees will not have the option to select it. Essentially, RF decorrelates the trees and therefore makes the results more reliable (Kuhn, 2013). The implementation of RF chosen only had one tunable parameter: the number of randomly selected predictors available for each tree split.

Support Vector Machines (SVM) were first proposed by (Cortes and Vapnik, 1995). This family of classifiers are the most complex models tested in this study. SVMs have two notable features: firstly, they are inherently binary classifiers, and secondly, they create linear hyperplanes (Murphy, 2022; Kuhn, 2013). SVM initiates by identifying datapoints of opposing classes proximal to one another. It then attempts to find the hyperplane that is equidistant from both sets of points, known as the maximum margin hyperplane. The opposing datapoints used to create this hyperplane are referred to as the support vectors. By this definition, SVMs are linear classifiers, but have been adapted to create non-linear decision boundaries. Three SVM models are used in this study; linear SVM, SVM with radial basis function, and SVM with polynomial kernel. The linear SVM only has one tunable parameter: the cost. The radial SVM has two parameters: sigma and cost. The polynomial SVM has three parameters: the degree, scale, and cost.

Hyperparameter Optimization

For models which did not have any tunable hyperparameters, such as MLR, LDA, and QDA, the training was conducted using 10-fold cross-validation repeated 10 times with accuracy used as the metric. For models which contained tunable hyperparameters, a grid search technique was employed for each of the hyperparameters, with 10-fold cross-validation repeated 10 times applied to each set of hyperparameters. Hyperparameters for each model (if present), and the selected values are presented in Table S2. The hyperparameter optimization curves for each of the models are shown in Fig. S5.

Feature Importance using Shap Values

Shapley values were used to test the hypothesis regarding the importance of aspect ratio with respect to the other shape features in supervised ML models. Shapley values were first introduced by Lloyd Shapley in 1953 (Shapley, 1953) to fairly distribute winnings between players based on their contribution to the game. The two pillars of Shapley values are additivity, where the sum of the winnings of each player must equal the total winnings, and fairness, where the highest performers cannot receive a lower share than the lowest performers. A concise explanation of the mechanism is as follows; in a scenario containing 4 players, in order to identify the importance of Player 1, all possible subsets of the players are made with and without Player 1. In the subset containing Player 1, the amount Player 1 receives is calculated, and in the subset without Player 1, the other players share Player 1's winnings. The difference between the amounts of both subsets gives the marginal contribution of Player 1, and therefore the overall importance of Player 1. Shapley values were theoretically proven as the fairest possible manner to distribute winnings. Lundberg et al. (2017) appropriated this concept from cooperative game theory into artificial intelligence (AI) to impute the importance of input features within black-box models (a field now known as 'Explainable AI'). To differentiate from its usage in game theory, the authors coined the term Shap values. Some models such as MLR and Random Forest have built-in variable importance measures. For MLR it is the magnitude of the coefficient, whereas Random Forest computes variable importance from the mean decrease in Gini impurity at each split of the decision trees, as well as the mean decrease in overall out-of-bag accuracy. However, most models do not provide this information and are effectively black boxes. Shap values are advantageous in that they are model-agnostic and retroactive with respect to the model building process, offering an external check used to explain the feature contributions to the predictions. It is important to note that Shap values calculate the local importance of features, which is the importance of a particular feature to a specific subset of datapoints. An aggregation is performed to provide the global importance of each feature with regards to the entire dataset.

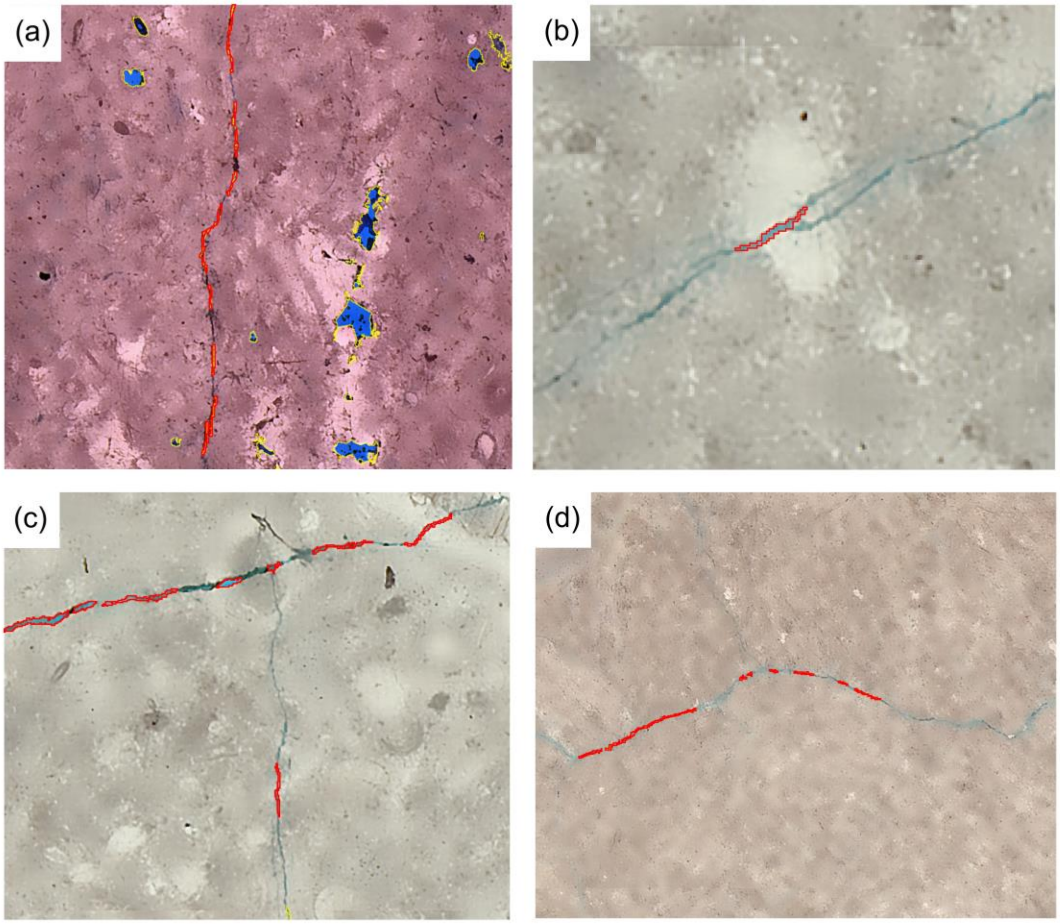

Figure S1. Fragmentation of microfractures from the thin section images. The red outlines indicate the segmented portions of the microfracture.

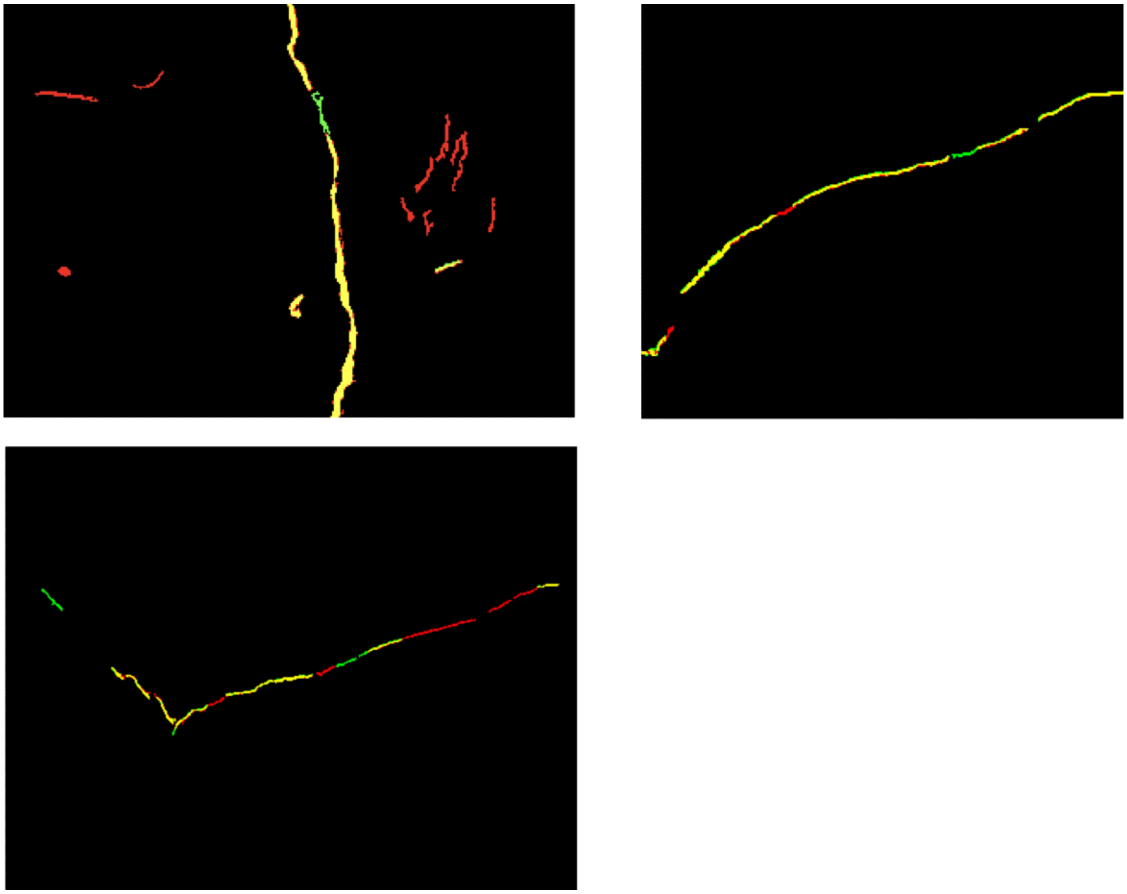

Figure S2. Composites of the HSB and LAB binary masks. Red signifies the HSB binary mask, green is the LAB binary mask, and yellow is the union of both masks. The HSB mask displays a stronger segmentation overall, but the LAB mask provides a notable boost in connectivity.

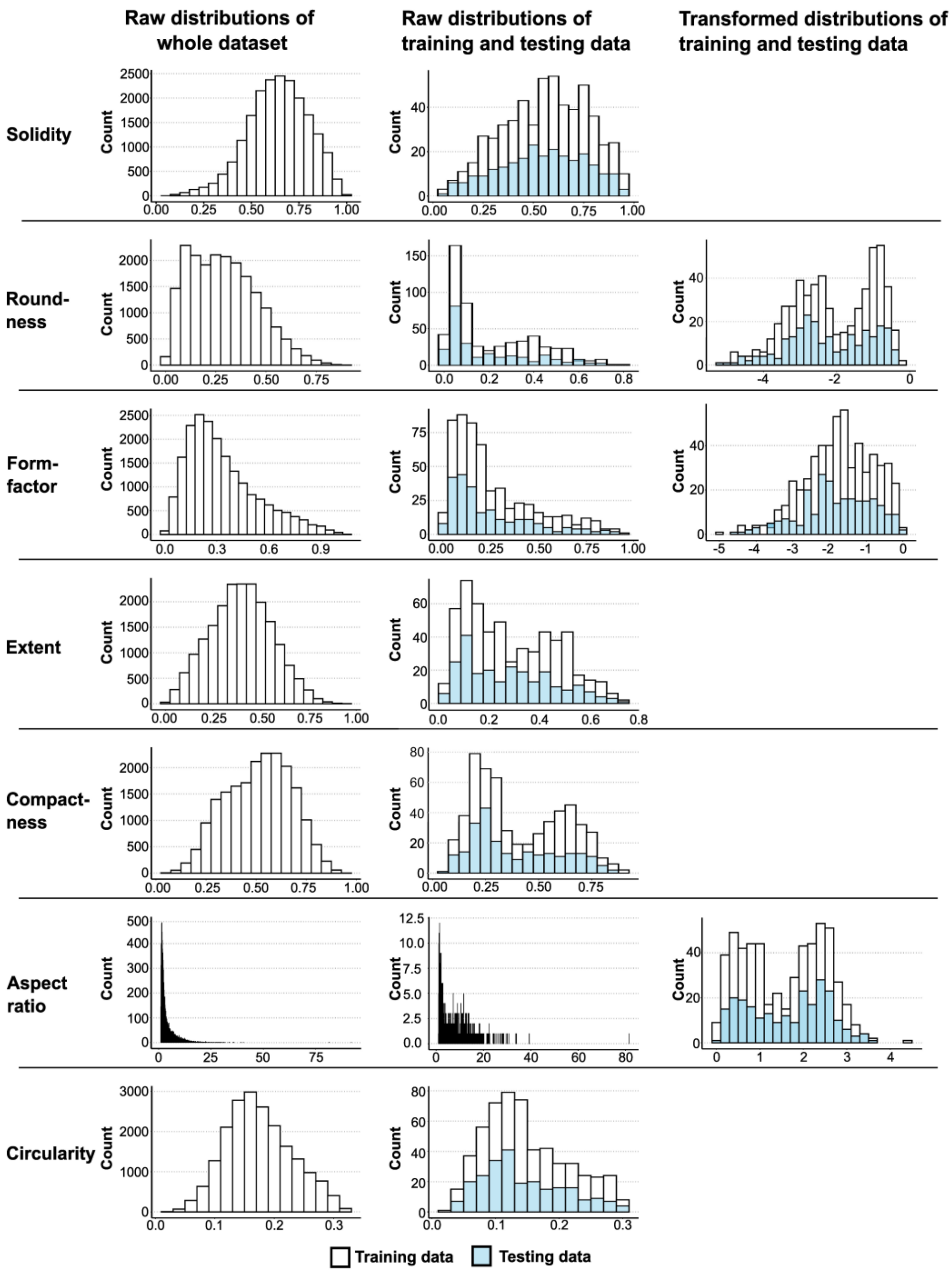

Figure S3. Univariate distributions of the shape features for the raw global dataset, the raw labelled dataset, and the transformed labelled dataset.

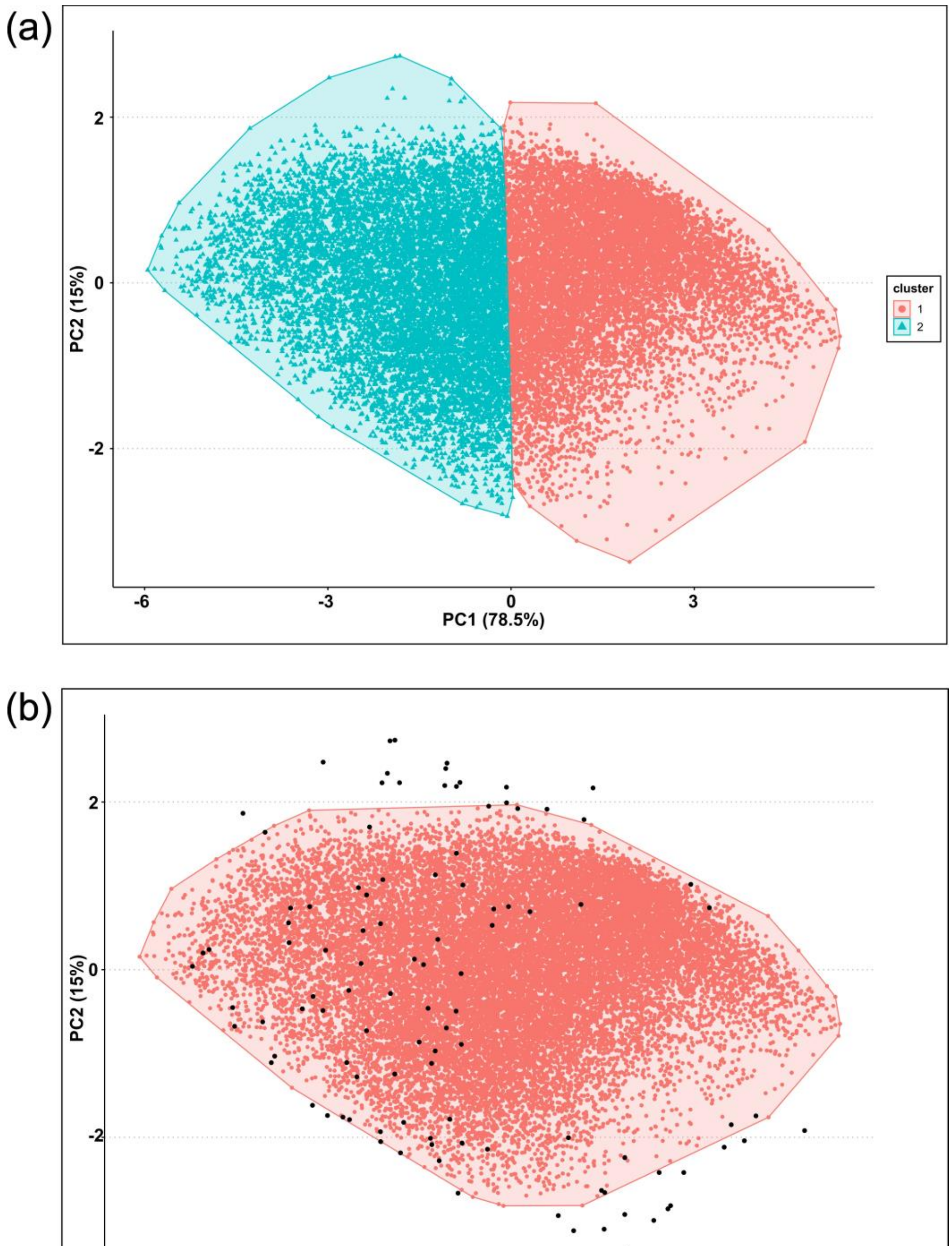

Figure S4. (a) Result of k-means on the global dataset. (b) Result of DBSCAN on the global dataset.

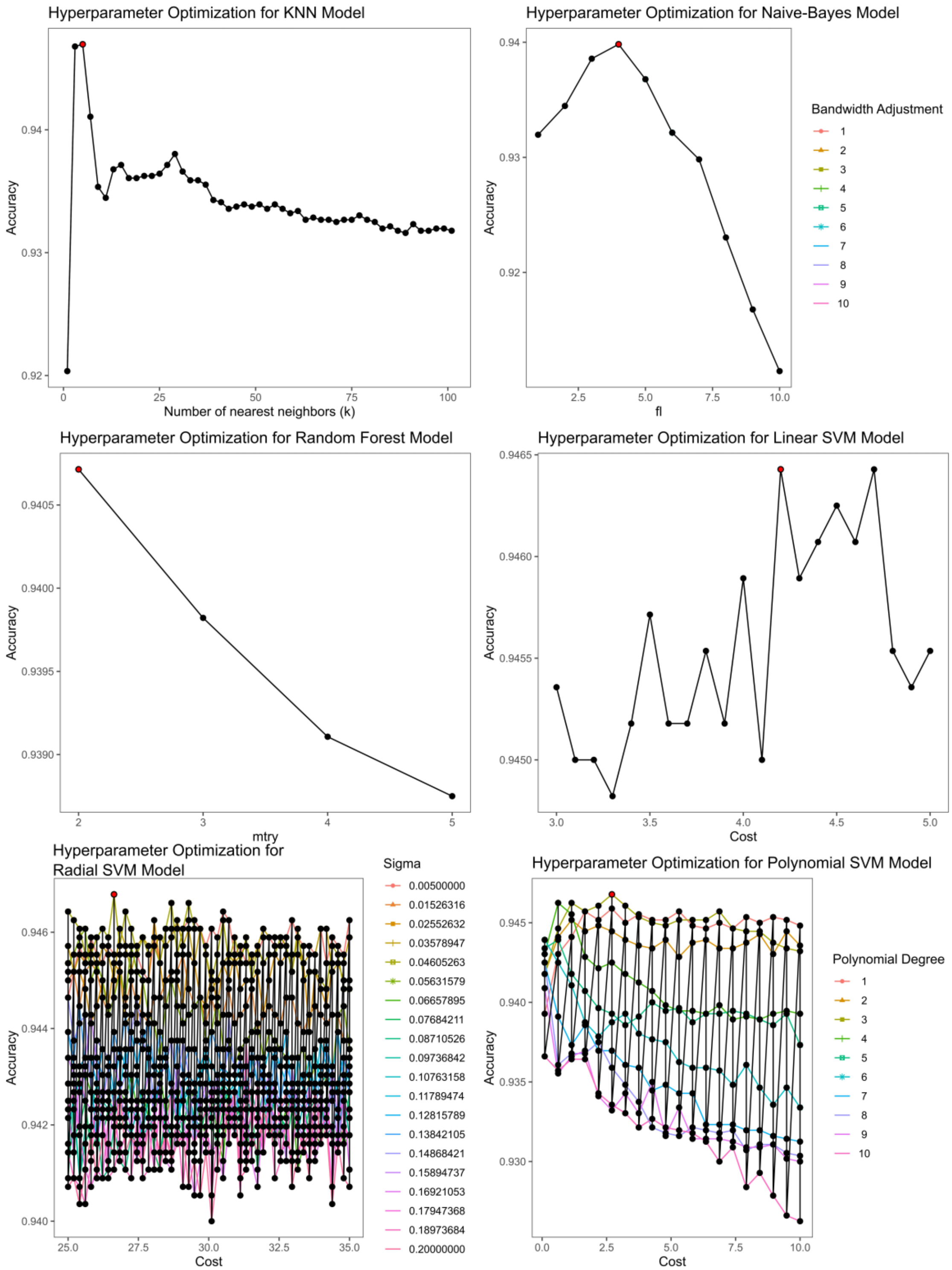

Figure S5. Hyperparameter optimization visualizations for the supervised ML models. MLR, LDA, and QDA did not contain any tunable hyperparameters and hence not included.

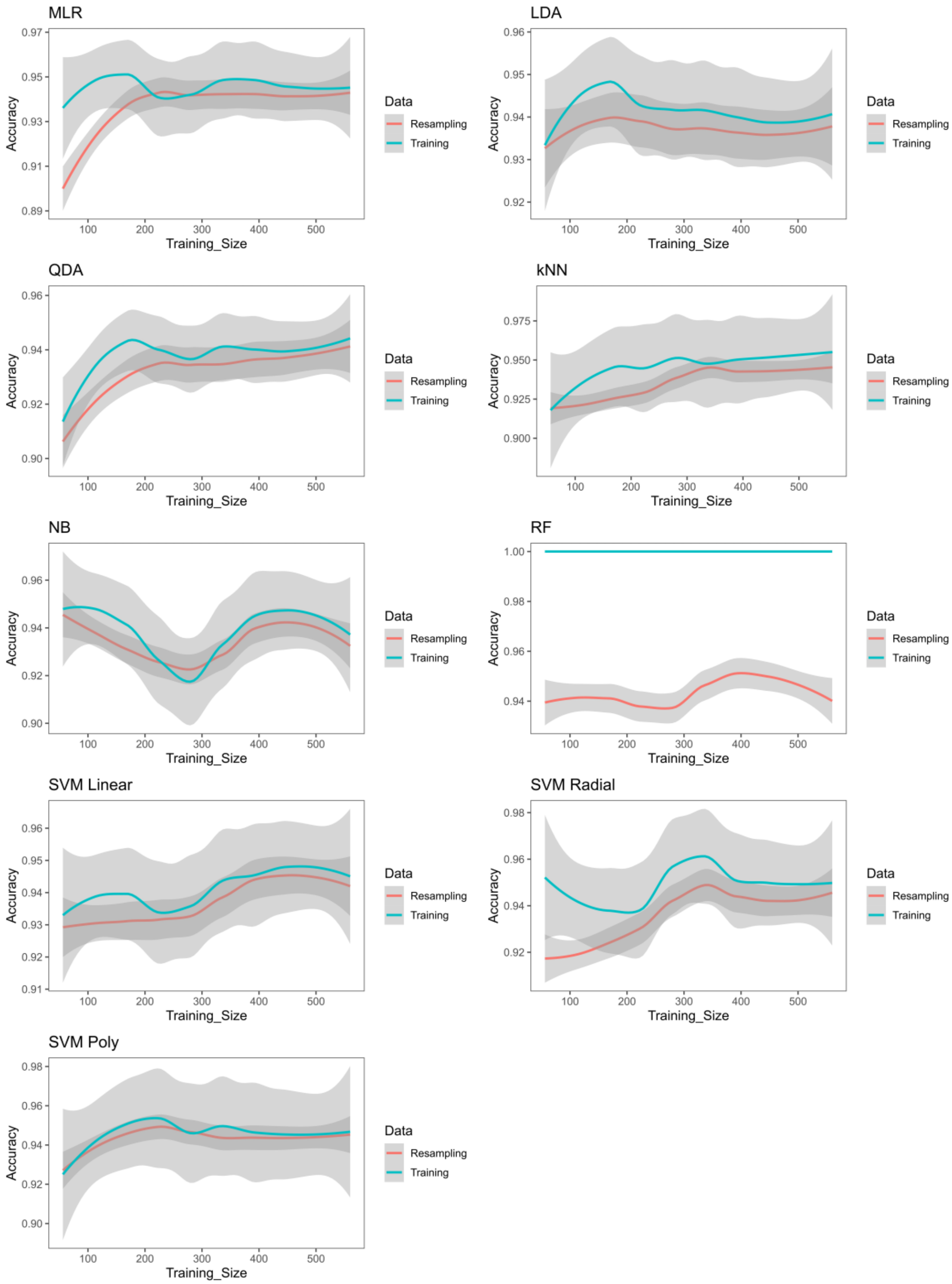

Figure S6. Learning curves for all the supervised models. Random forest was the only model which showed overfitting as the training accuracy was constantly 100% with the resampling accuracy significantly lower.

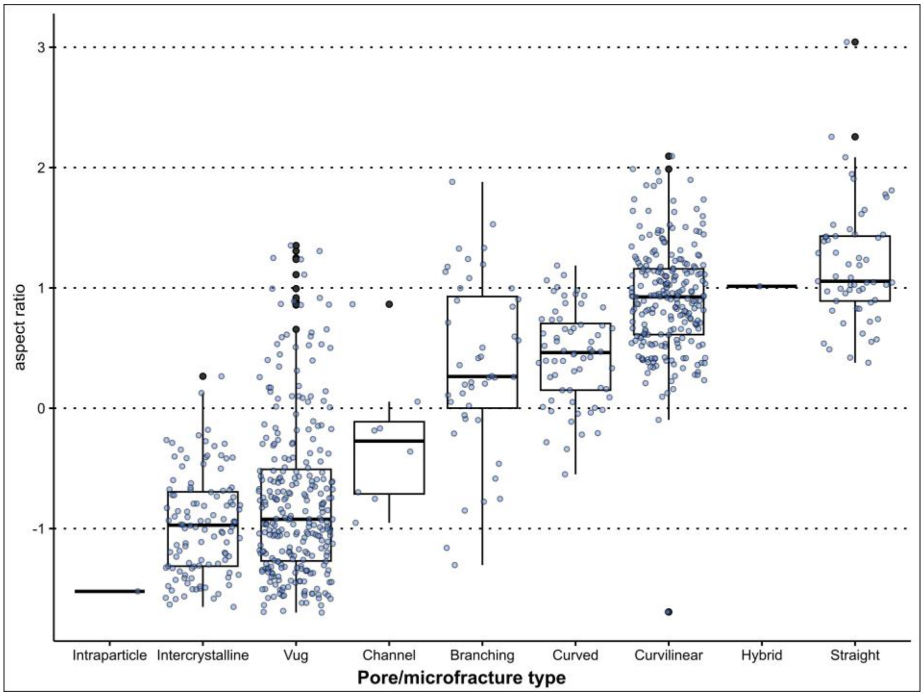

Figure S7a. Boxplot of aspect ratio ranges for the secondary labels of pore and microfracture types.

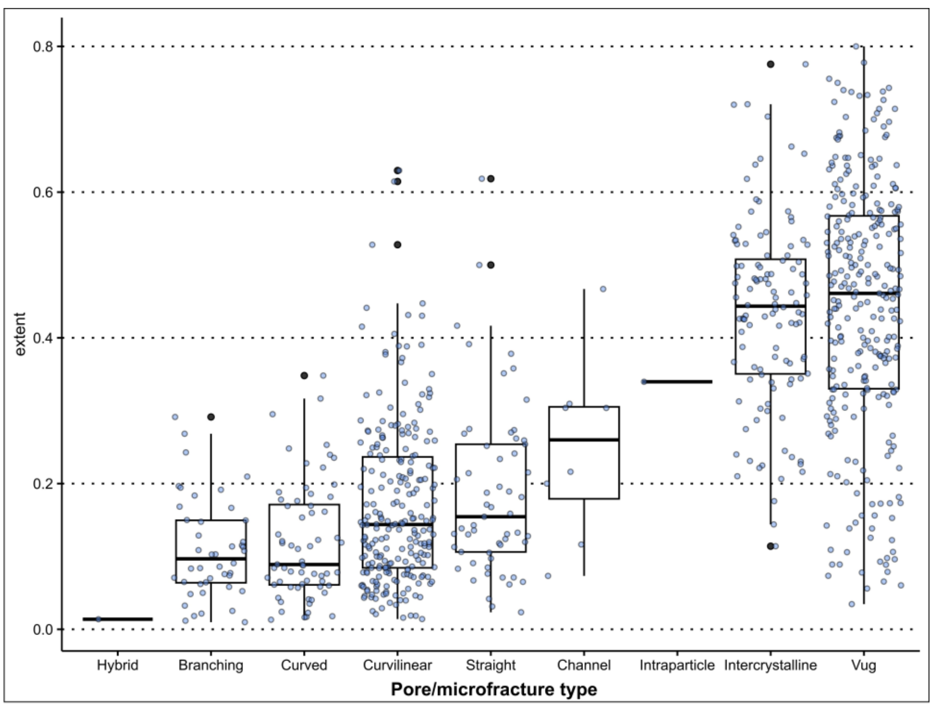

Figure S7b. Boxplot of extent ranges for the secondary labels of pore and microfracture types.

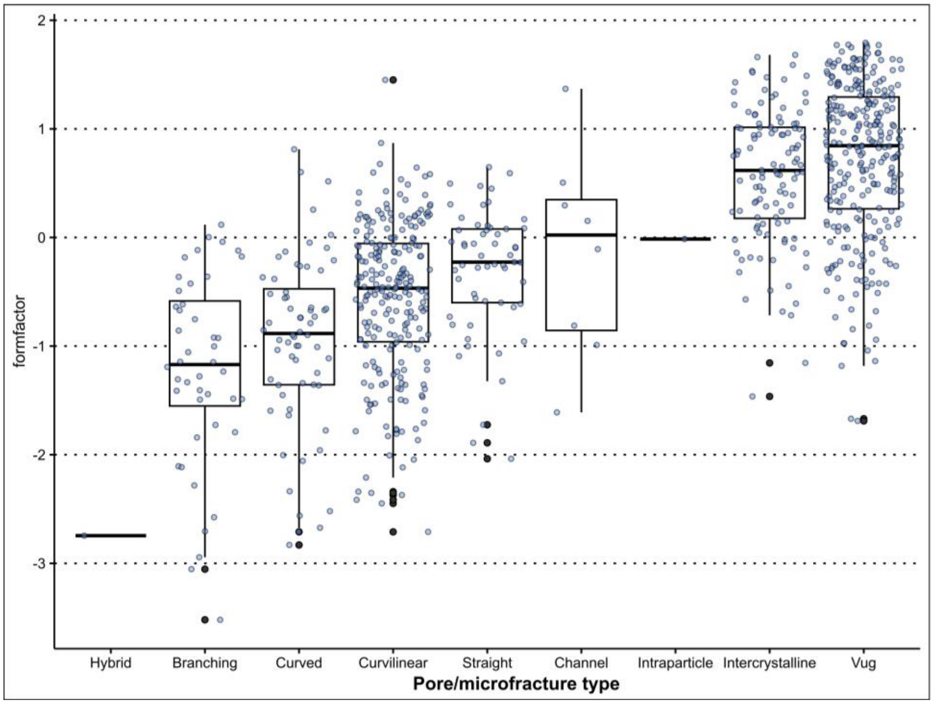

Figure S7c. Boxplot of formfactor ranges for the secondary labels of pore and microfracture types.

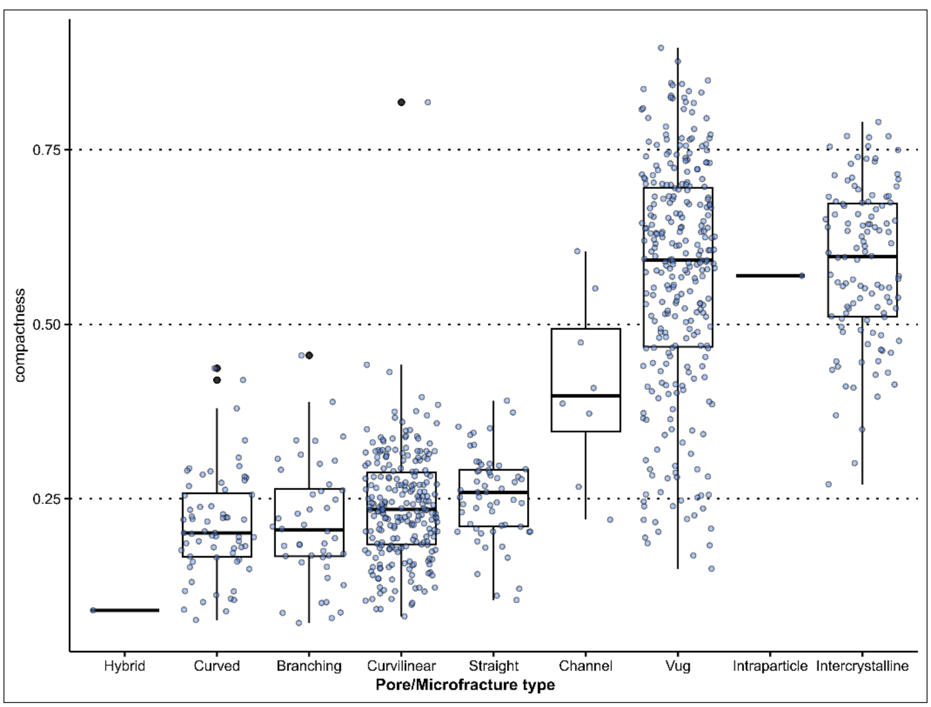

Figure S7d. Boxplot of compactness ranges for the secondary labels of pore and microfracture types.

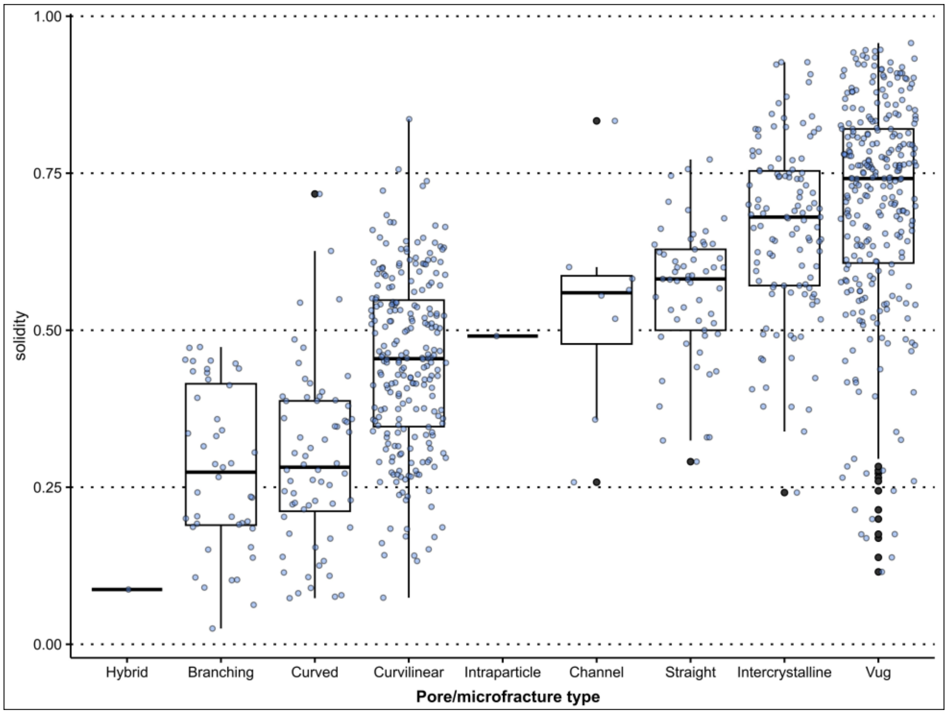

Figure S7e. Boxplot of solidity ranges for the secondary labels of pore and microfracture types.

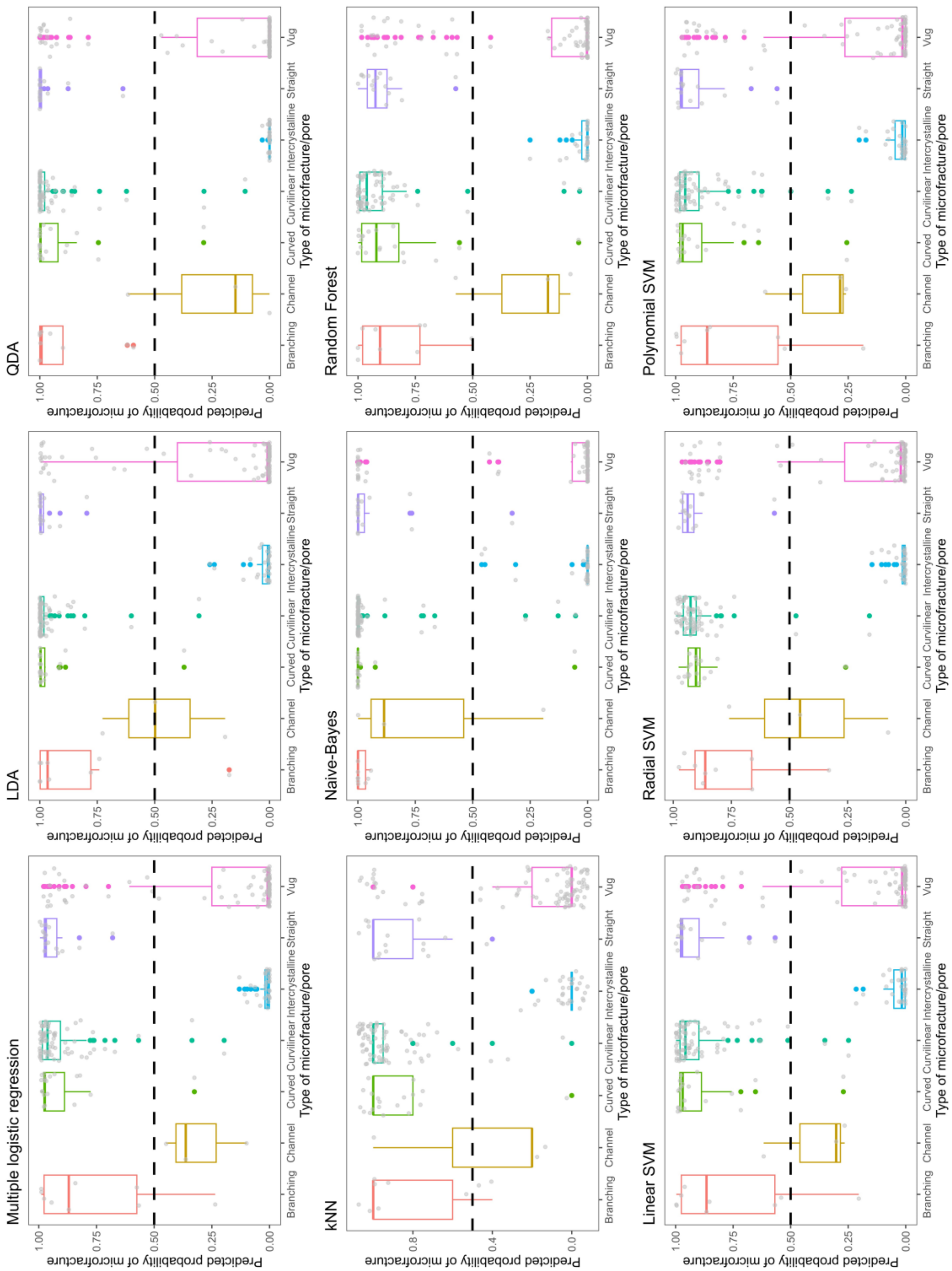

Figure S8. Probability of microfracture prediction per secondary label for each supervised model. The dashed line represents the 50% decision threshold, any objects above 50% are classified as microfractures and any object below the threshold are classified as pores.

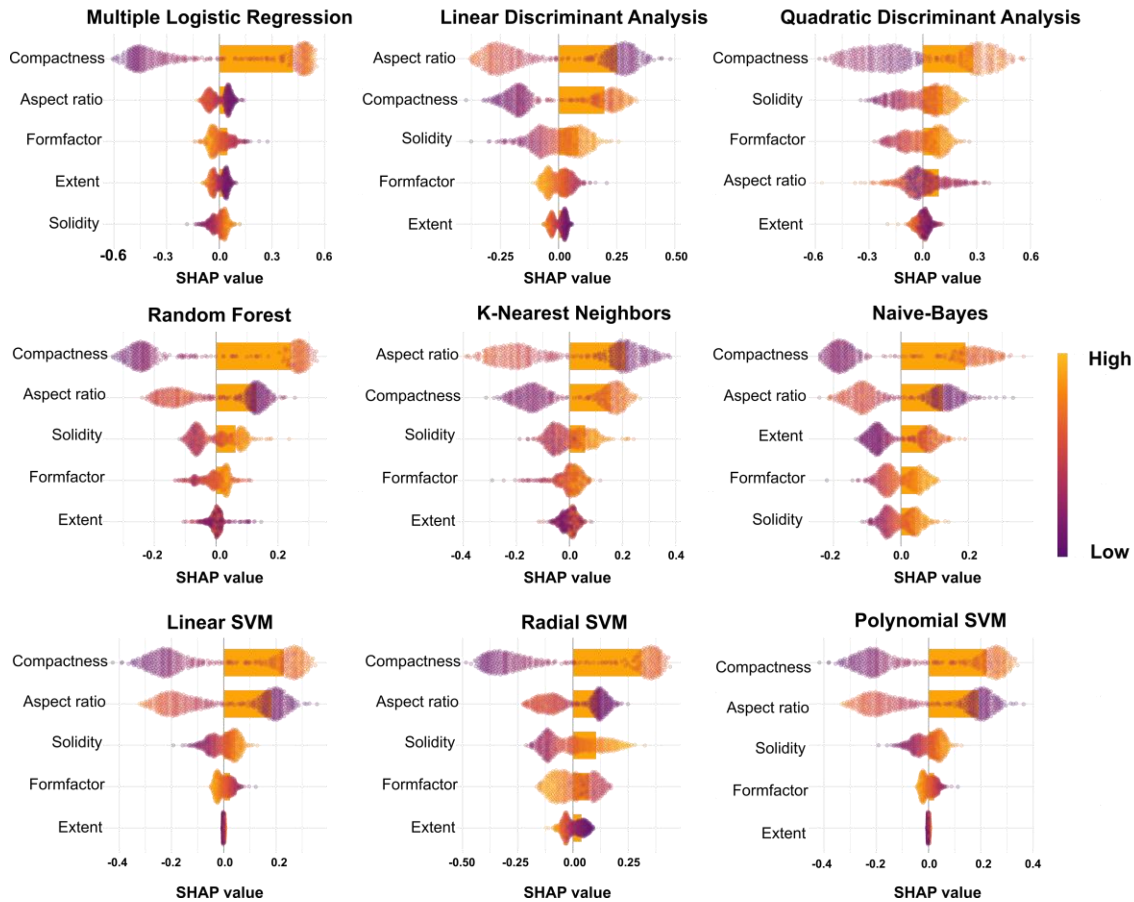

Figure S9. Probability of microfracture prediction per secondary label for each supervised model. The dashed line represents the 50% decision threshold, any objects above 50% are classified as microfractures and any object below the threshold are classified as pores.

Table S1. Studies on automated pore typing using AI.

Table S2. List of hyperparameters for each supervised ML model and the final values chosen.

Model	**Number of hyperparameters**	**Hyperparameters**	**Final values**
MLR	0	-	-
LDA	0	-	-
QDA	0	-	-
kNN	1	Number of neighbours (k)	k = 3
Naive-Bayes	3	Laplace (fL), Kernel, bandwidth adjust (BA)	fL = 2, Kernel = True, BA = 2
Random Forest	1	Number of randomly selected variables at each split (mtry)	mtry = 2
Linear SVM	1	Cost (C)	C = 0.3
Radial SVM	2	Cost (C) and Sigma	C = 22.63, Sigma = 0.04
Polynomial SVM	3	Cost (C), degree of polynomial, and scale	C = 0.974 , degree = 4, scale = 0.1

Data Set S1. Image data for the study.