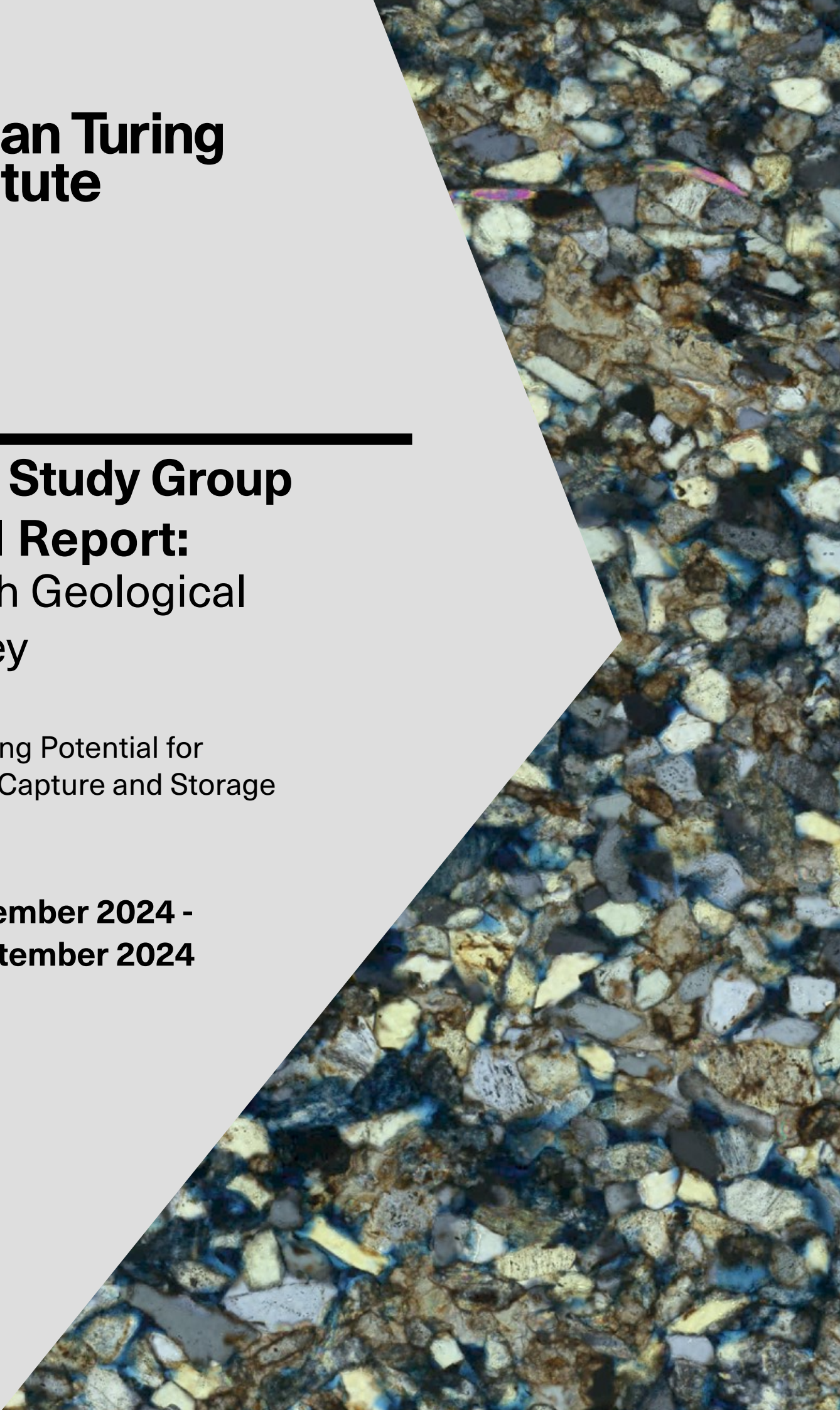


The Alan Turing Institute

Data Study Group Final Report: British Geological Survey

Identifying Potential for
Carbon Capture and Storage
in Rock

**9 September 2024 -
13 September 2024**



Contents

1	Executive Summary	3
2	Introduction	5
2.1	Background	5
2.2	DSG Objectives	6
3	Data overview	7
3.1	Dataset description	7
3.2	Data exploration and quality	9
4	Approach	11
4.1	Research Questions	11
5	Porosity mapping	12
5.1	Results	14
5.2	Limitations	18
5.3	Future work	19
6	Grain segmentation	19
6.1	Results	21
6.2	Limitations	26
6.3	Future Work	26
7	Mineral Classification	28
7.1	Random Forest	28
7.1.1	Data Preprocessing	29
7.1.2	Model Training	30
7.1.3	Results	32
7.1.4	Limitations	34
7.1.5	Future Work	39
7.2	U-Net	40
7.2.1	Data Preprocessing	40
7.2.2	Model Training	41
7.2.3	Results	42
7.2.4	Limitations	44
7.2.5	Future work	45

8	Discussion	46
9	Conclusion	48
10	Team members	49
	References	51
	Appendix	55

1 Executive Summary

The injection and sequestration of CO₂ in deep geological repositories is a promising strategy for reducing atmospheric CO₂ concentrations. Deep storage of hydrogen could act as a ‘battery’ for storing energy long term. Both technologies are supporting efforts to achieve the targets established by the UK Government’s Net Zero energy strategy [11]. Deep repository injection is a costly and complex undertaking; therefore, a careful, advanced assessment of the suitability of a candidate rock formation for injection and storage is essential and urgently needed. The suitability of a candidate rock formation depends on several properties of its sedimentology microstructure, and conventional analysis methods are time-intensive and costly on a per-sample basis. The development of an automated and reliable characterisation process would be highly beneficial in determining the suitability of rock formations for injection and storage.

The objective of this Data Study Group Challenge was to develop and demonstrate automated methods for the characterisation of microscopic rock samples using images of thin-section samples provided by the British Geological Survey. The dataset consisted of 119 thin-section rock samples, each imaged under eight different lighting conditions, 13 of which had additional labelled images. The labels indicated the mineral grains within the images. We divided the challenge into three related tasks: assessing porosity (using pixel-wise thresholding rules), identifying grain boundaries (using the watershed algorithm), and determining mineralogy (using the random forest and U-Net algorithms).

We demonstrated that these approaches show promise for automating thin-section analyses and highlighted opportunities for leveraging the results of each task to improve the others. Porosity estimation can serve as an initial step for the watershed algorithm in grain detection, which, in turn, can define grains for applying the random forest feature-based mineral classification. The identified grain boundaries can be used to refine the mineral label images, which were available at a lower resolution than the other images. These refined label images could then be used to train the U-Net algorithm for mineral detection.

We have identified several potential areas for further work, as each task could undergo iterations to enhance the algorithms. A key area for improvement across all tasks is the incorporation of a greater number of input image layers (i.e., different lighting and polarisation conditions) to refine the outputs. Due to time and

resource constraints, not all eight available lighting condition images were utilised for each task. Learning from all images simultaneously could enhance porosity estimation, grain boundary detection, and mineral classification. Additionally, this challenge facilitated discussions regarding the explainability of the algorithms used. The random forest feature-detection approach follows a more traditional human analytical process by identifying grains and their optical properties to perform classification. This approach may be preferable in some applications compared to the U-Net classification method, which utilises the optical input data in a less interpretable manner to make mineral classification decisions.

Images in this report contain data from the British Geological Survey © UKRI 2024.

2 Introduction

2.1 Background

Climate change is an existential challenge for humanity in the 21st century, driven by anthropogenic sources of greenhouse gas emissions (GHGs) [15]. A rapid increase in atmospheric concentrations of GHGs since the Industrial Revolution has led to a rise in global average temperatures and an increasing frequency of extreme weather events, as the Earth's climate system undergoes unprecedented changes [18]. The broad scientific consensus indicates that, without substantial mitigation efforts, these changes will accelerate, resulting in severe consequences for ecosystems, human health, and socio-economic systems [15, 20].

Increases in atmospheric CO₂ concentrations have been a primary driver of anthropogenic climate change, with levels currently exceeding 420 parts per million (ppm)—the highest in at least 800,000 years [18], with anthropogenic carbon sources driving this rapid increase [15]. Anthropogenic CO₂ amplifies the natural greenhouse effect and is a dominant contributor to climate change, with CO₂ emissions from human activities accounting for roughly three-quarters of total global greenhouse gas emissions [28].

One strategy for reducing atmospheric CO₂ levels is carbon sequestration in geological formations, where CO₂ is injected into deep underground rock formations. This process, often referred to as Carbon Capture and Storage (CCS), targets depleted oil and gas fields as well as deep saline aquifers as repositories for the sequestration of captured CO₂. Within suitable geological formations, CO₂ may be stored for thousands of years without significant risk of leakage [27]. The British Geological Survey (BGS) has identified over 500 potentially suitable geological sites within the UK for large-scale CO₂ storage [9].

The storage of hydrogen fuel in geological formations has also been proposed as a viable solution for large-scale energy storage to complement intermittent renewable energy sources. Hydrogen fuel would be produced when there is an abundance of renewable energy, injected into geological storage, and withdrawn as needed [9].

A series of prospecting studies conducted since the mid-1990s have identified numerous depleted oil and gas fields and deep saline aquifers in the UK North Sea as potentially viable formations for CO₂ sequestration and hydrogen fuel storage [14, 13]. A national database (CO₂Stored) has been developed to catalogue potential

storage repositories [9], allowing for the assessment of geological formations' suitability for sequestration and storage, thereby supporting the UK's Net Zero energy strategy.

Thin rock slices, commonly referred to as thin sections, are used to characterise porosity and mineralogy in potential storage reservoirs. These slices are produced from rock cores extracted from formations considered for sequestration or storage and are cut to a thickness on the micrometre scale.

Digital imaging from transmitted visible light microscopy is used to examine the microstructure of rocks and determine properties such as sample porosity, grain size, and distribution. These features impact a formation's ability to transmit and store fluids. The mineral composition of the rock affects geochemical reactions such as dissolution, precipitation, and the liberation of fine particles; thus, understanding the composition is essential for assessing formation suitability for gas injection.

Thin section analysis provides critical data for determining the suitability of geological formations for carbon or hydrogen injection, particularly in relation to long-term storage potential and reservoir integrity. However, current thin section analysis is a time- and labour-intensive process, making manual interpretation impractical for large sample sets. The BGS holds an archive of over 500,000 thin section samples, necessitating an automated process for characterising key parameters as the only feasible approach.

2.2 DSG Objectives

The objectives of the BGS-DSG challenge are:

- (i) The development of models to identify the structure (grains and pores) of thin section rock samples
- (ii) The development of models to identify the composition (types of grains) of thin section rock samples
- (iii) The development of an analytical pipeline to automatically analyse sample images and deploy the methods for structure and composition analysis

Quantities of interest to characterise from thin-slice samples include:

- Porosity (the amount of space between mineral grains)

- Mineral composition (which minerals are present, and in what proportions)
- Grain sizes and shape
- Directionality of non-spherical grains
- Surface textures
- Distributions of the numerical parameters among samples

These quantities summarise the geology of the sample and inform decision-makers about the suitability of a site for gas storage. Porosity is a critical factor, as the gas requires both sufficient space for storage and channels for injection. Low porosity would indicate a poor candidate for gas injection and storage.

The presence of minerals that react with gas can be beneficial for the long-term sequestration of CO₂; however, for hydrogen storage—where the storage is temporary—such minerals would be contraindicators of a site’s suitability. Other grain features are also important in assessing how easily the injected material can flow into the available spaces. All of this information is contained within the thin-section samples.

3 Data overview

3.1 Dataset description

The data of interest in this study consist of a collection of sandstone thin-section samples extracted from the Bunter Sandstone Formation [26], located in the UK Southern North Sea. The locations are shown in Figure 1. Many sandstones are naturally porous, making sandstone formations viable candidates for gas injection and storage.

The dataset comprises imaged versions of physical rock samples provided by the BGS. The thin-section samples were scanned using a machine that exposes the sample to different lighting conditions, producing eight aligned RGB images. Whilst the physical samples vary in age, with some collected as early as the 1980s, the scanned images were generated recently using consistent methodology and the same equipment. These images are used by geologists to determine the mineralogy of the rock. They are extremely high resolution, with dimensions on the order of 10^4 pixels per axis.

A limited number of samples have undergone expensive electron microscope scanning, including both backscatter electron microscopy (BSE) and scanning electron microscopy (SEM/phasemap). The phasemap data are of significantly lower resolution than the RGB images but provide valuable information about the texture (BSE) and mineral composition (phasemap) of the samples. These data can be used as labelled training data. The phasemap and BSE data were manually aligned with the RGB image data as part of pre-processing before being provided for the Data Study Group (DSG), and artefacts may have been introduced during this process.

The phasemap data consist of RGB images in which colours represent 18 different mineral and artefact labels, whereas the BSE images are greyscale, with intensity values corresponding to the intensity of electrons bouncing back from the thin-section surface to the scanner.

In total, there are 119 samples, each with eight images captured under different lighting conditions (including circular light, polarised light, and cross-polarised light at multiple angles). Of these, 13 samples have associated labelled data (phasemap and BSE).

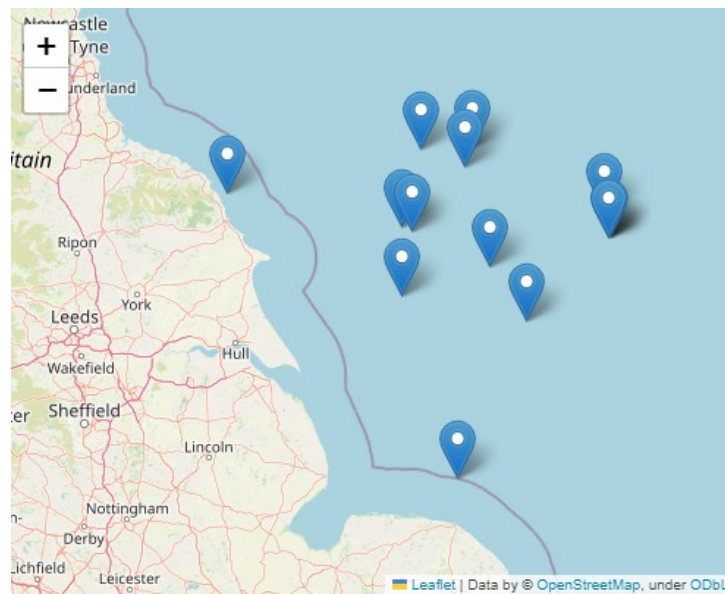


Figure 1: Locations of where the rock samples within this study were taken.

3.2 Data exploration and quality

For each sample, the group of images created under different lighting conditions are of the same size and are aligned. The labelled data (phasemaps) have a lower resolution compared to the thin-section images; however, they contain the same number of pixels and have been manually aligned with the image data (by BGS before the DSG). Addressing the differences in resolution through down-sampling or super-resolution techniques was considered to be outside the scope of the DSG, so the analysis does not account for these resolution differences. Within each thin-section sample, the group of images are of the same size; however, across different samples, the images may vary in size.

We demonstrate the dataset using the first sample from the training set. These images have 25923×45990 pixels. Figure 2(a) shows a standard light condition image. The sample does not fill the image and is not square to the edges of the image. The phasemap for this sample is shown in Figure 2(b). The phasemaps are aligned to the image data, however they do not span the extent of the samples as the phasemaps are created using a separate analysis technique and have only been generated for sub-sections within the sample. The rest of the images in the figure show the same 512×512 tile for each of the images. Figure 2(c) is the colour image (CPOL) and Figure 2(d) shows the sample in plane polarised light (PPL). Figures 2(e) - 2(j) show the sample in cross polarised light at different angles (XPL). Figure 2(k) shows the same section of the sample analysed using a backscatter electron microscope (BSE) as a greyscale image. Figure 2(l) demonstrates the phasemap for the subsection, this is of a lower resolution so appears pixelated. The phasemap is in the format of an RGB image, however the phasemap RGB pixels take only 18 distinct colour values representing different mineral classes.

The thin section is imaged using a scanner that produces optical images (such as Figures 2(c) - 2(j)), so these layers (CPOL, PPL, XPLs) are available for each sample in the dataset. The BSE and phasemaps are only available for a limited number of samples in the dataset. The grains show up as different colours in the different cross polarised images (Figures 2(e) - 2(j)). The boundary between the large two grains in the sample is visible in many of the optical images, but not in the BSE (Figure 2(k)) or phasemap (Figure 2(l)) images. In this dataset, there are no grain boundary labels available for comparison. The pore space appears light blue in the PPL image (Figure 2(d)) but is difficult to discern in the other optical images. The pores show as black space in both the BSE (Figure 2(k)) and phasemap (Figure 2(l)) images.

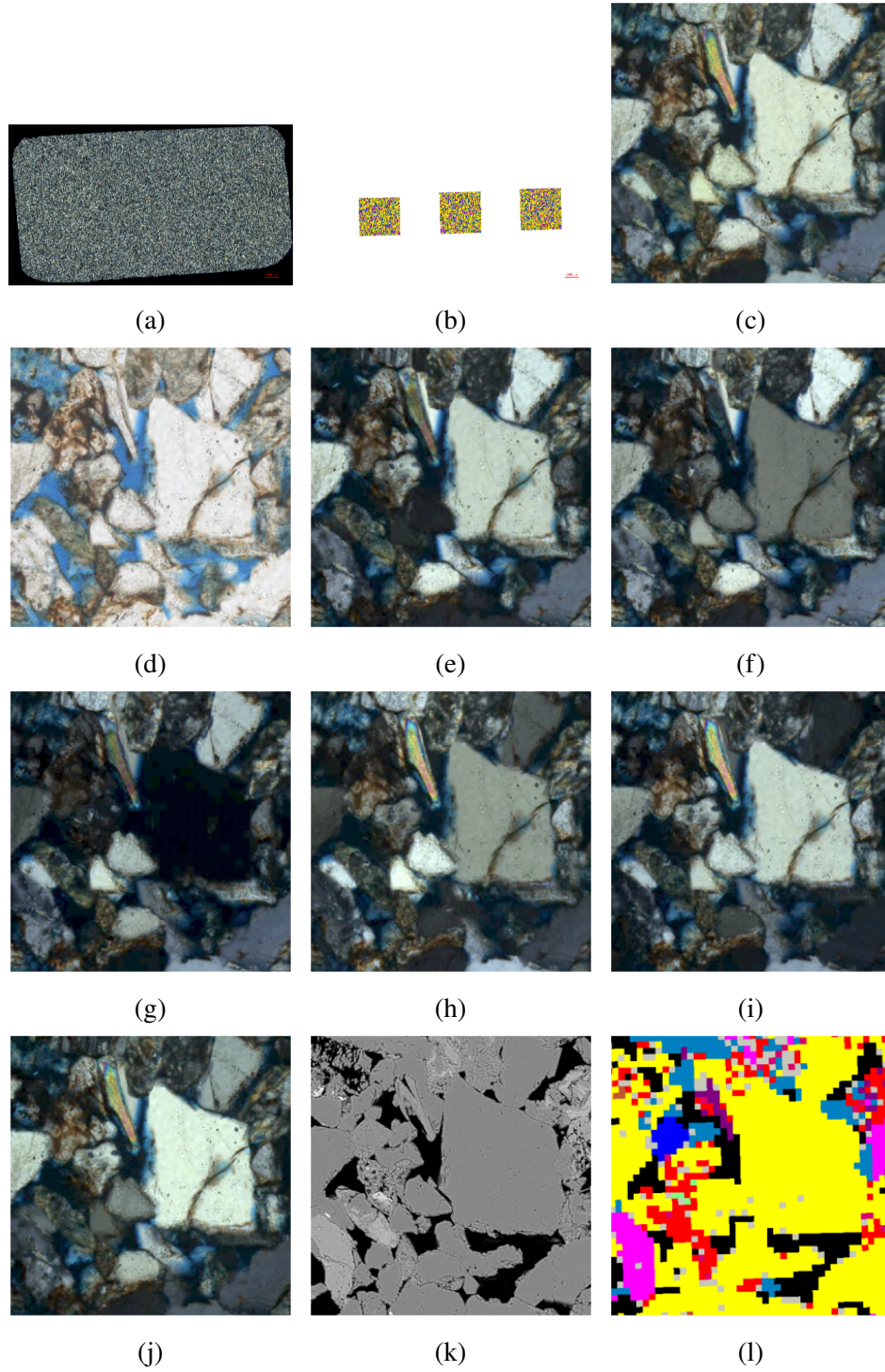


Figure 2: First sample images from the training set. (a) Full CPOL image (b) Full phasemap (c) Tile from CPOL (d) Tile from PPL (e) Tile from XPL0 (f) Tile from XPL15 (g) Tile from XPL30 (h) Tile from XPL45 (i) Tile from XPL60 (j) Tile from XPL75 (k) Tile from BSE (l) Tile from phasemap

The quality of the provided data is discussed as follows:

- **Appropriateness:** The data are aligned with the questions of interest as they are typical thin section images for a sandstone geological formation which is a potential storage location. The mineral phasemaps provide information about mineral composition for some samples which were available as labelled training data.
- **Readiness:** Additional metadata such as the labels for phasemap images and the locations of the samples were available. In addition, a table of mineral classes and examples were provided for additional context about the different light condition images. The images were very large so required pre-processing and tiling before performing analyses. The phasemaps were in RGB rather than class label form, so these had to be converted using a dictionary before use.
- **Reliability and Bias:** The data represent thin section samples from a particular geological formation, so could not be considered representative of all thin section sample images. In addition, all of the images were generated from the same scanning machine so could be subject to unknown machine distortions. These data can be used to consider the questions of interest for sandstone thin sections.
- **Sensitivity:** The data are scientific images of rock segments so there is no personal information or privacy concerns associated with these data.
- **Sufficiency:** Whilst many sample images were available, there were few labelled images. Although we were able to work with the existing labelled data, additional labelled data would enhance future work on this project. Higher-resolution phasemap data would also be beneficial, although this may not be feasible due to instrument resolution limitations. Additionally, labelled grain boundary data would enable the investigation of a broader range of techniques for grain boundary detection.

4 Approach

4.1 Research Questions

To accomplish the objectives described in Section 2.2, the following work packages were set up by the team:

- (i) Estimation of porosity values of thin slices using raw images and thresholding rules. This is a binary classification task, separating the images into grains and pores.
- (ii) Detection of individual grains from raw images. This is an edge detection task.
- (iii) Classification of the minerals based using mineral phasemaps as training data. This is a multi-class semantic segmentation task. Our team has assessed two different approaches to this classification:
 - A deep-learning algorithm that uses raw images as inputs;
 - A random forest feature-based classification algorithm that will use features of the colour histograms of grains as input.

The diagram of the image processing workflow is presented in Figure 3.

These tasks were worked on in parallel during the DSG. However, they would be deployed sequentially in a data processing pipeline. Using the masks created from these tasks the properties of the rock can be quantified. The tasks are interrelated; porespace provides the background for identifying grain boundaries. Grain boundaries identify individual grains for random forest analysis. Finally, the CNN output also includes porosity information.

5 Porosity mapping

The following method focuses on determining the porosity of rock samples using the PPL (Plane Polarised Light) images provided within the data set. Pores in PPL images of rock samples appear as light blue because the pores are filled with resin during sample preparation. This resin fills in the empty spaces (pores) within the rock, and under visible light, the resin appears as light blue. When observed in PPL microscopy, these light blue areas indicate where the resin has filled in the pore spaces, making the pores visible for analysis. Black regions in the PPL images are typically created by oxide minerals, which are not part of the pore space. Due to the large size of the samples, the PPL images are divided into smaller, manageable tiles for processing.

Colour thresholding isolates the light blue pixels from the rest of the image. The method used here applies the following RGB colour thresholds that correspond to

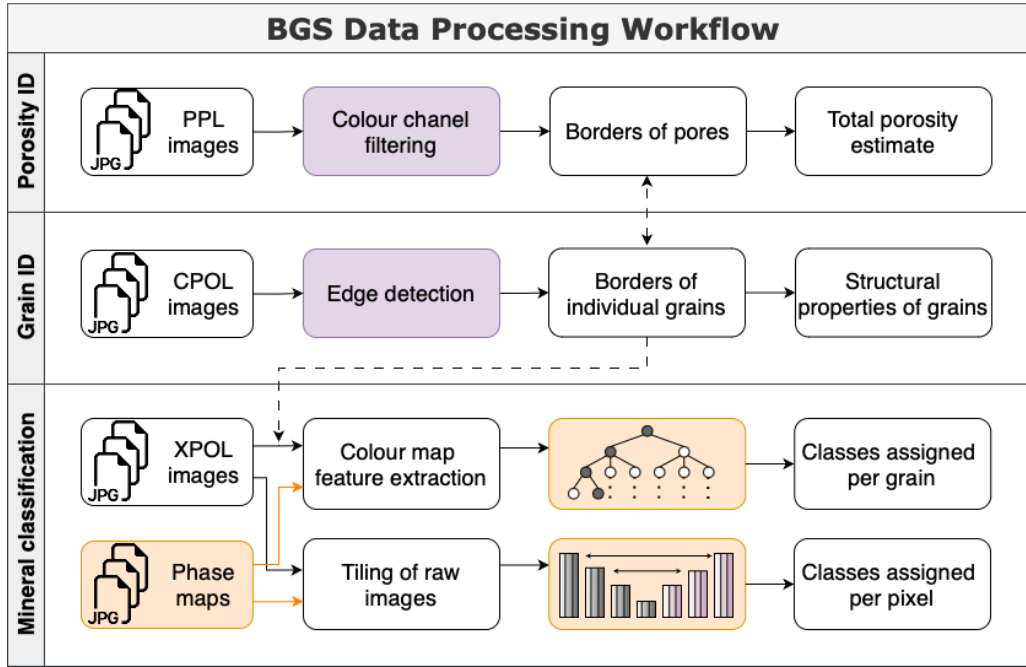


Figure 3: Schematic of data processing workflows used for the characterisation of thin section sandstone samples

the light blue hue, separating the pore spaces from the rest of the rock matrix:

- **Red Channel:** values between 0 and 150
- **Blue Channel:** values between 50 and 220
- **Green Channel:** values between 130 and 255

Once the light blue regions are isolated, the percentage of light blue pixels is calculated relative to the total number of pixels in the image or tile. This percentage represents the porosity of the sample (i.e. the volume fraction of the pores relative to the total rock volume). This method was evaluated within the DSG as it offers a computationally inexpensive solution to estimating porosity.

For the training dataset, the porosity determined from PPL images is compared with BSE (back-scattered electron) images and phasemaps using sections where there is phasemap information available. This allows for comparison of the porosity estimates, as pores in both the BSE and phasemap appear black on the images. Applying similar colour thresholding approaches generates a similar

binary classification per pixel. Values below 50 for all RGB bands are classed as pores for BSE images and exactly 0 for all bands for phasemap. An estimate of porosity for the training data is also given in the BGS report [26], which is based on the percentage of black pixels across the entire phasemap image. For the broader dataset, which does not include BSE or phasemap data, the same PPL method is applied to estimate porosity based solely on the light blue proportions in the images. The results are compiled into histograms to show the distribution of porosity across multiple samples, shown in Section 5.1.

5.1 Results

In this section, we explore percentage porosity calculation through the colour mapping of PPL images. Using the training dataset, we compare PPL-derived porosity values with those from the BSE and phasemap images, using 512×512 and 1024×1024 pixel images. Porosity in PPL images is determined by isolating light blue regions corresponding to resin-filled pores, while in BSE and phasemap images, it is calculated by identifying black regions representing pore spaces. This comparison allows us to evaluate the accuracy of PPL-based porosity estimates and validate them against more established methods. The results are shown in Table 1.

Table 1: Percentage porosity estimated using different image types and sizes within a training sample

Tile Size	Image Type	% Porosity
512×512	PPL	17.76
	BSE	15.44
	Phasemap	13.44
1024×1024	PPL	15.79
	BSE	13.24
	Phasemap	11.54

The BGS report [26] indicates a porosity value of 13.6% for this sample, a value calculated from the full phasemap image of the entire sample (rather than a tile). This explains its close alignment to the phasemap percentage porosity of the training dataset images. The percentage porosity estimated using the PPL images

is greater than when estimated with BSE images, with the phasemap images having the lowest percentage. Phasemaps are generated at a lower resolution than the other images, which may explain the lower porosity as small areas of porespace will not be captured at the coarser resolution. Percentage porosity is lower for larger image sizes, which is not representative of the whole sample as only two image tiles were compared here. However, the relative percentage porosity calculations are similar for both sizes. The PPL, BSE and phasemap for the 512×512 training data image before and after colour mapping are shown in Figure 4.

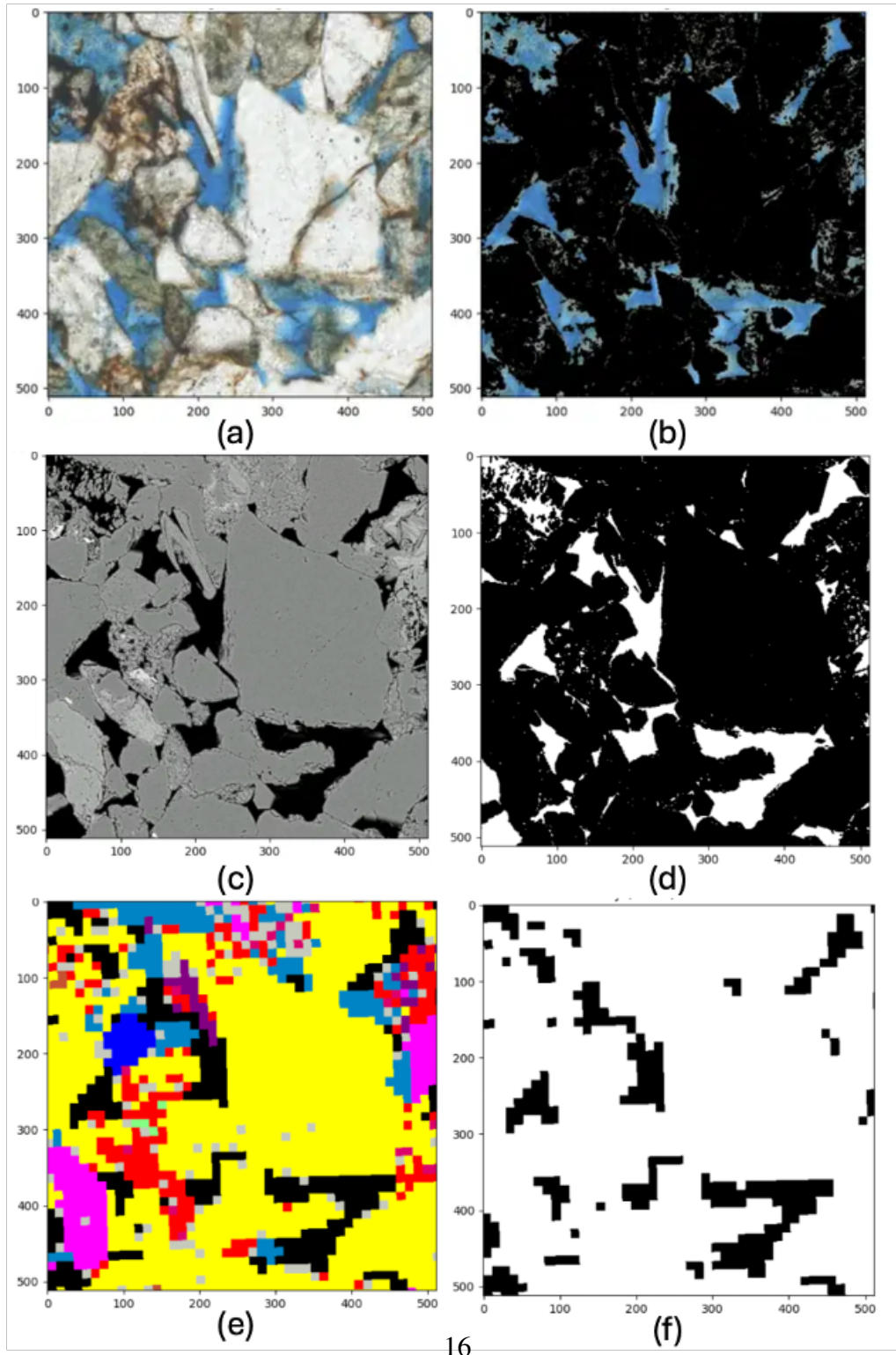


Figure 4: Porosity comparison on a training dataset image tile of size 512×512 (a) PPL image; (b) PPL with pore colour mapping (blue marks pores); (c) BSE image; (d) BSE with colour mapping (white marks pores); (e) Phasemap image; (f) Phasemap with colour mapping (black marks pores)

Applying PPL image colour mapping to the wider 119 image data set, and taking a larger 6122×5632 tile size (the same dimensions as the area of interest in the first training data image), gives a range of percentage porosities shown in the histogram in Figure 5. The range of porosity expected from this type of geology is 13-30%, however some samples fall outside of this range. This indicates either issues with the area in the images that was selected for evaluation (as the samples do not fill the images) or unexpected features of the samples.

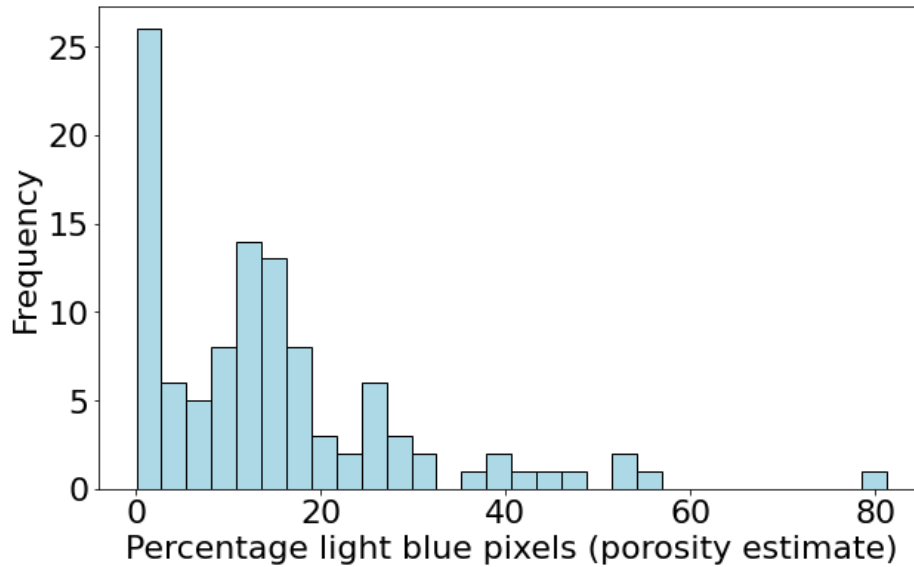


Figure 5: Histogram of estimated porosity across all samples using the PPL images

Of the 119 sample images, 56 were taken from one location, well number 42-25d-3, all at varying depths. Figure 6(a) shows the relationship of depth and porosity at this location. Figure 6(b) shows the depth vs porosity plot at an alternative location, well number 44-23a-10, where 11 of the samples were taken from. The first scatter indicates a wide variation in the porosity of samples, which could be down to a high variation in rock porosity vertically, or occasional poor sample quality towards smaller depths. The second scatter plot shows a slightly clearer trend in results, with porosity increasing with depth. The value above 80% porosity is an artefact as porosity values of this magnitude are physically impossible.

The porosity estimation can be used as part of a longer pipeline of image analysis tasks and is an efficient method to evaluate the porosity across large images due to

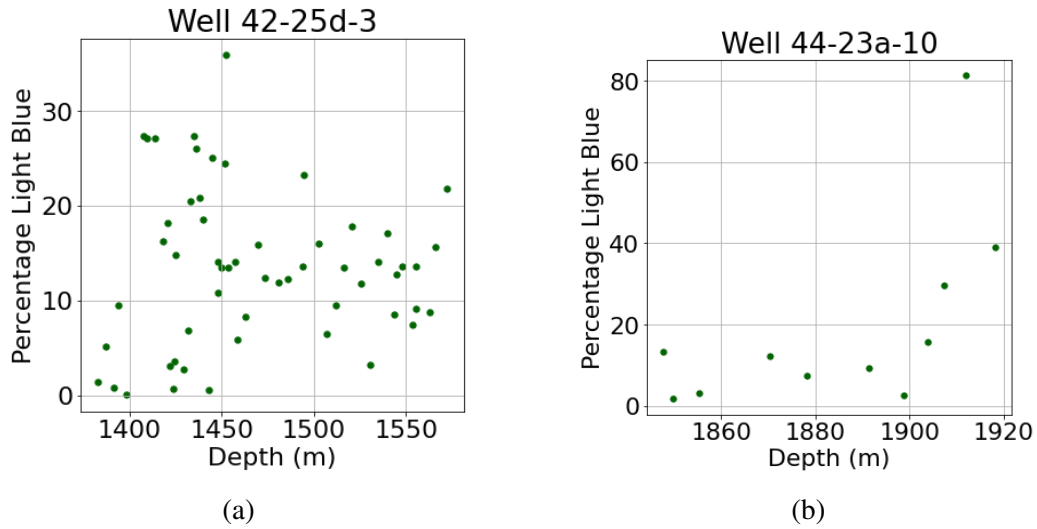


Figure 6: Estimated porosity against depth for well (a) 42-26d-3 (b) 44-23a-10 (using PPL images)

its rule-based method.

5.2 Limitations

This approach to porosity estimation is a thresholding check of colour, so has some limitations.

- **Data:** The provided data did not include maps that label the grains and pores at the same resolution as the images. This makes it difficult to benchmark the porosity estimation against ground truth values as the only available labels were the phasemap images at much lower resolution, or BSE images where porosity again had to be estimated with a thresholding approach.
- **Approach:** There may be subtle features that this binary mapping does not handle, such as tiny slivers of grains with resin over or under them. This approach also relies on the use of blue resin when preparing the samples, but alternatives are sometimes used.

5.3 Future work

Further work to hone this porosity estimation technique would be improved by more detailed labelled data. Conversion to a production-ready algorithm would include the following steps:

- Creation of a tool to identify the edges of the samples so that the porosity estimate can be bounded by these (non square) edges.
- Further evaluation of the porosity estimation over different sample sizes to identify whether sub-samples of the data could be used for more efficient porosity estimation.
- Checks could be added to the porosity estimation to indicate unusual samples or issues in the analysis, such as proportion of black or white pixels in the analysed region. These checks could add a flag to the porosity estimates to indicate those that should be looked at in more detail, or not trusted as realistic estimates.

6 Grain segmentation

In this section, we describe investigations conducted to identify the boundaries and regions of individual grains. As there is no labelled data available for grain edges, this constitutes an unsupervised task. We apply the watershed algorithm to various images, exploring different parameters to segment grains and visually assessing the results. The watershed algorithm has been successfully employed for similar grain detection problems [23].

The watershed algorithm treats an image as a topological landscape defined by intensity levels, such as greyscale values, where the intensity determines the altitude of the landscape at each point in the image. Basins in the landscape are considered the starting points for watersheds, which are filled with water at a rate dependent on the gradient of the landscape, as illustrated in Figure 7. When a watershed reaches a point where it begins to merge with another, a boundary is established, representing the division between two objects (watersheds) in the image. The final set of boundaries delineates the grain segmentation.

The watershed algorithm also separates the background (in this case, the pores) from the foreground (grains). It is possible to define a ‘sure background’ at the start of the watershed algorithm using binary thresholding of the image. This

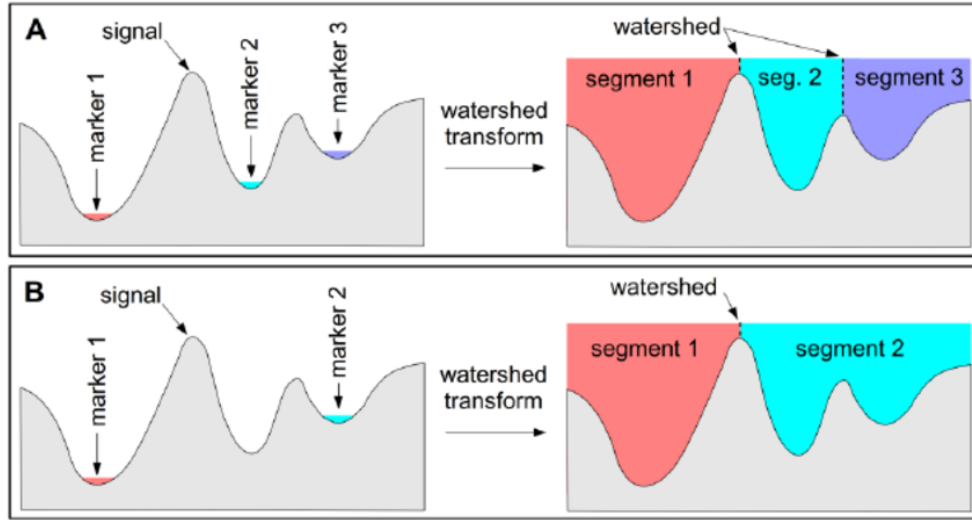


Figure 7: Diagram representing the watershed algorithm for in one dimension. Figure 3 (unaltered) from [10], ©A. Fisher, licence CC-BY 3.0

ensures that the designated background region is guaranteed to be part of the pore space in the final segmentation.

The initial basins selected (markers in the image) each correspond to an individual grain in the output, making the determination of the number and position of these markers highly important.

We apply the watershed algorithm using the OpenCV module in Python [5] to both CPOL (circular polarised) and BSE (back-scattered electron) images, investigating various image preprocessing techniques to improve performance.

For method exploration, two square regions of size 512×512 pixels were selected from the first sample to evaluate the watershed approach. These regions were chosen as they exhibited significant differences in porosity—one being highly porous and the other considerably less so. CPOL, PPL, XPL, and BSE images were examined as potential inputs for the watershed algorithm. These are shown in Figure 8.

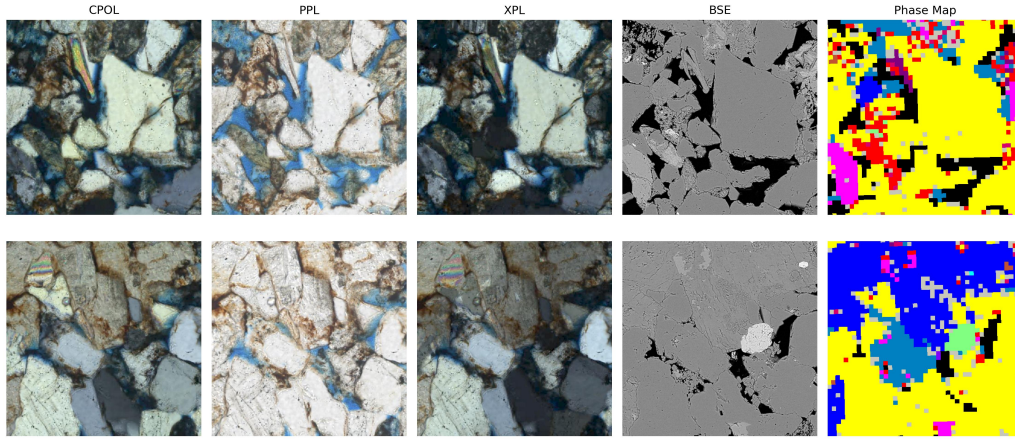


Figure 8: Two chosen tiles (top row: more porous, bottom row: less porous), imaged in circular polarised (CPOL), plane polarised (PPL), cross polarised (XPOL) light, backscatter electron (BSE) and phasemap forms

6.1 Results

The watershed method was first applied to the CPOL images for the two image subsets, using a standard approach (main tutorial from `OpenCV`¹). This is visualised in Figure 9. The steps are as follows:

1. The CPOL image (1st image in Figure 9) is transformed to greyscale and then to binary image using a greyscale threshold that is determined by the Otsu algorithm [21](an automatic method that determines the threshold based on the histogram of greyscale intensities). Dark greyscale values are assigned to black and light to white.
2. Noise from the image is removed using a morphological operation to repeatedly erode and dilate the binary regions.
3. Dilation of the white region is applied to slightly reduce the the black background region. The remaining black region is set as the ‘sure background’ and will remain as the background when the watershed is applied (2nd image in Figure 9).
4. The remaining white region is the potential foreground, i.e. grains. A

¹https://docs.opencv.org/4.x/d3/db4/tutorial_py_watershed.html

proportion of this is assigned as the ‘sure foreground’, while the rest is left as an ‘unknown region’ (3rd image in Figure 9). This proportion is determined by a distancing threshold, (set to 0.1 for this example). The unknown region will be assigned to either background or image segments.

5. Initial segment markers for watersheds are determined by the connected components of the ‘sure foreground’ (4th image in Figure 9).
6. Watershed algorithm is applied to the sure foreground and unknown region to assign to markers or background and get the final segmented image (5th image in Figure 9).

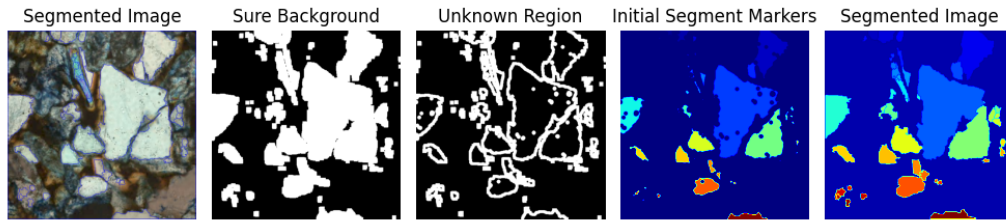


Figure 9: Watershed example steps: segmented image shows the input to the algorithm with the output superimposed in blue, sure background shows the background in black, unknown region indicates areas in the image that are categorised as neither foreground nor background during the processing, initial segment markers show in colour each of the connected regions within the sure foreground, and the segmented image shows the resulting watershed estimation applied to the markers and unknown region.

The initial thresholding is important for determining both the sure background and initial markers. Here, we compare a range of manually set thresholds (rather than the automatic Otsu threshold) to understand this relationship, as shown in Figure 10. It shows that the binary threshold level has a significant impact on the final segments and is optimal around 100 for this example, capturing the major grains while segmenting some connected features. Increasing the threshold above this level loses grains. For the lower threshold, more grains are captured in the foreground (as opposed to the dark blue background) but are less well-segmented. The performance of the best threshold setting is similar to results using the Otsu threshold so we continue with the Otsu method going forward as it has the benefit of being data-driven for each image.

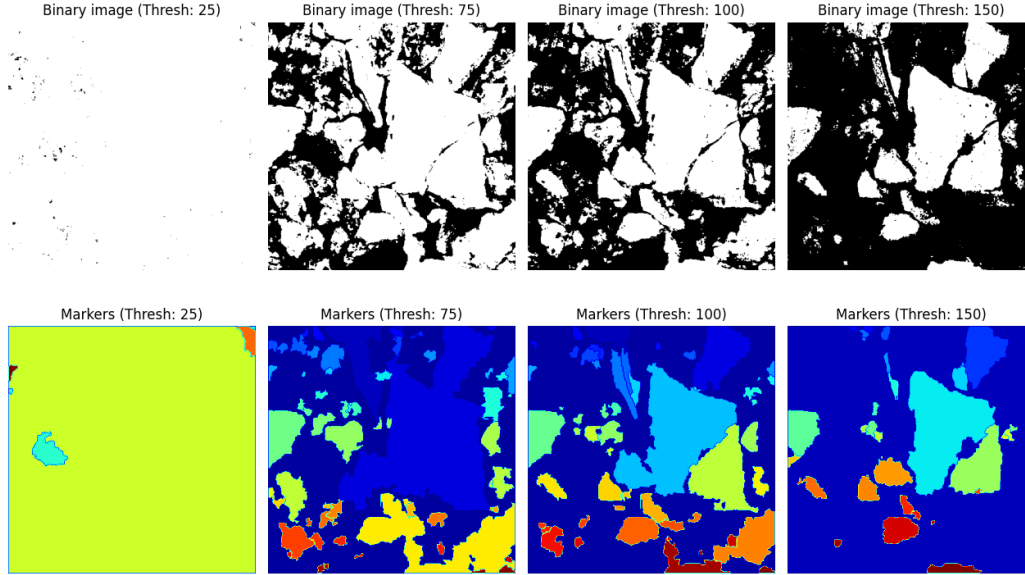


Figure 10: Comparison of binary thresholds. Top row: binarised image with different thresholds. Bottom row: resulting watershed markers.

Some large grain sections are identified that should have grain segment boundaries within them, so we preprocess the image to enhance edges between grains. Edges are enhanced using the Canny algorithm [7](which applies Gaussian filtering and thresholds edges based on gradients to highlight connected strong edges) and then reimposed onto the original image before applying the watershed algorithm. Results are shown in Figure 11. This results in oversegmentation of grains in places in the first image after edge enhancement. For the second image, the image is undersegmented for large grains without edge enhancement but then slightly oversegmented with edge enhancement as seen in Figure 11 (bottom left corner of lower right image).

We explored several additional improvements:

- Reducing the sure foreground area and increasing the unknown area after thresholding, based on the distance threshold value. We found that as this value increases, fewer grains are identified.
- Setting the sure background to match the black regions of the BSE image.

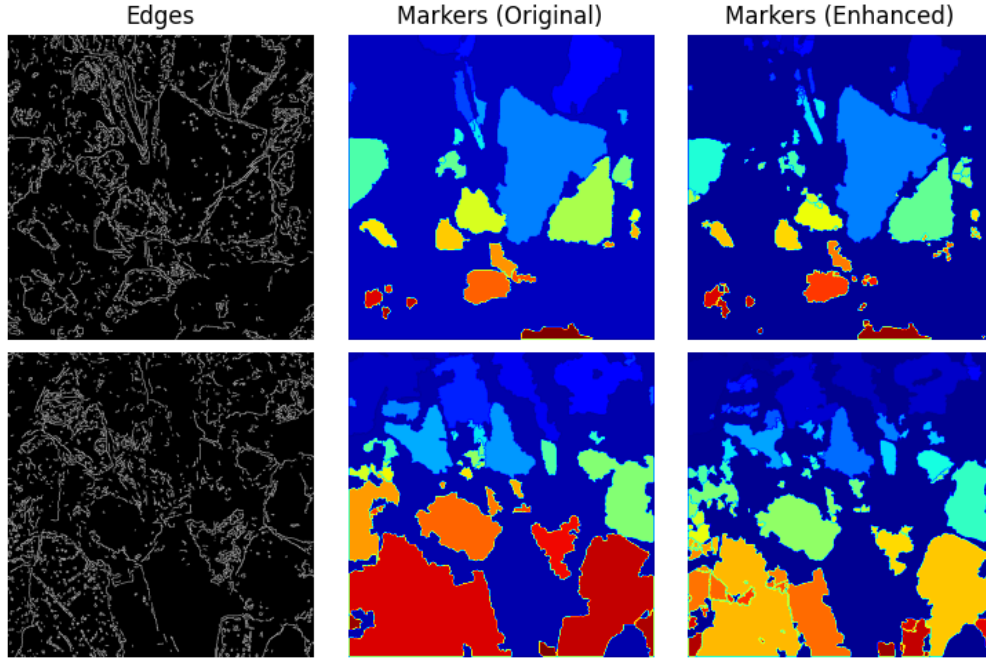


Figure 11: Comparison of markers with and without Canny edge enhancement. Left: edges detected with Canny edge detection, middle: markers for watershed algorithm without Canny enhancement, right: markers for watershed algorithm following Canny enhancement.

This, however, did not yield any performance improvements.

- Using different colour spaces (HSV: hue, saturation, and value; LAB: lightness, red-green, and blue-yellow), as shown in Figure 12. While the A and B channels of LAB exhibit lower contrast (making background separation more challenging), they have the potential to be preprocessed to capture some of the darker grains, as these would not be excluded during thresholding. We applied segmentation to the H, S, A, and B channels of the different colour spaces, with the results presented in Figure 13. The colour channels influence the selection of grains and the distinction between foreground and background. For instance, comparing the H and A channel results reveals that dark blue backgrounds and coloured grains appear in different regions of the segmentation outputs.

The challenge in achieving optimal watershed performance lies in both ensuring

that the background is correctly assigned and selecting initial markers from the sure foreground region in a way that generates the correct connected components.

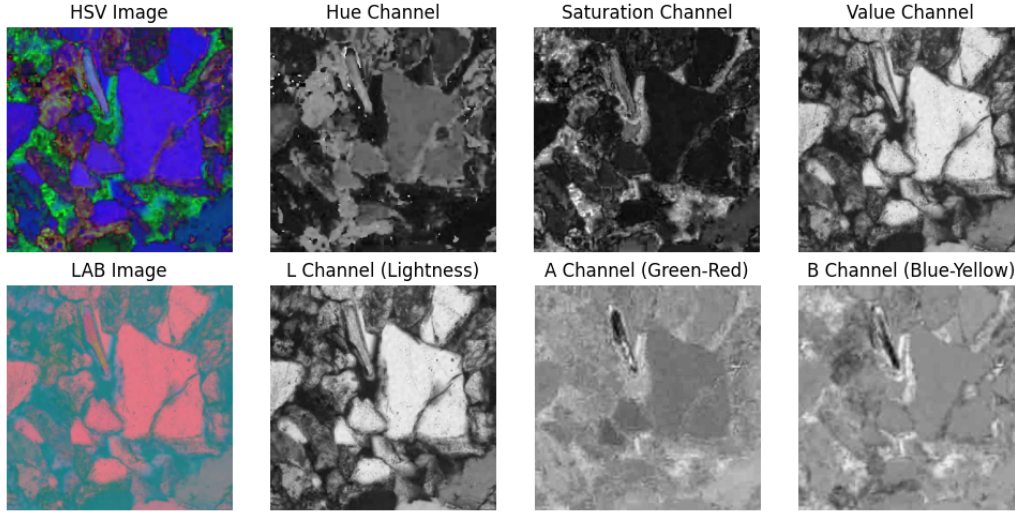


Figure 12: CPOL image shown in HSV (top row) and LAB (bottom row) colour spaces

We also consider the BSE image for segmentation, as it provides clear pore definition and some indication of grain boundaries. Figure 14 illustrates this approach with and without edge enhancement. In both cases, standard segmentation tends to under-segment, whereas the edge-enhanced version over-segments the grains. The optimal choice of algorithm depends on whether overestimating or underestimating grain boundaries is more detrimental to the analysis. The results using the BSE image appear to better capture smaller grains compared to those obtained from CPOL image segmentation.

Finally, the watershed algorithm is applied to a larger section of the BSE image, shown in Figure 15. This identifies 3862 distinct grains and 29.3% pore space. This is significantly higher than previous estimates of porosity, showing that many grains are missed from this segmentation and are instead considered as background or pore space.

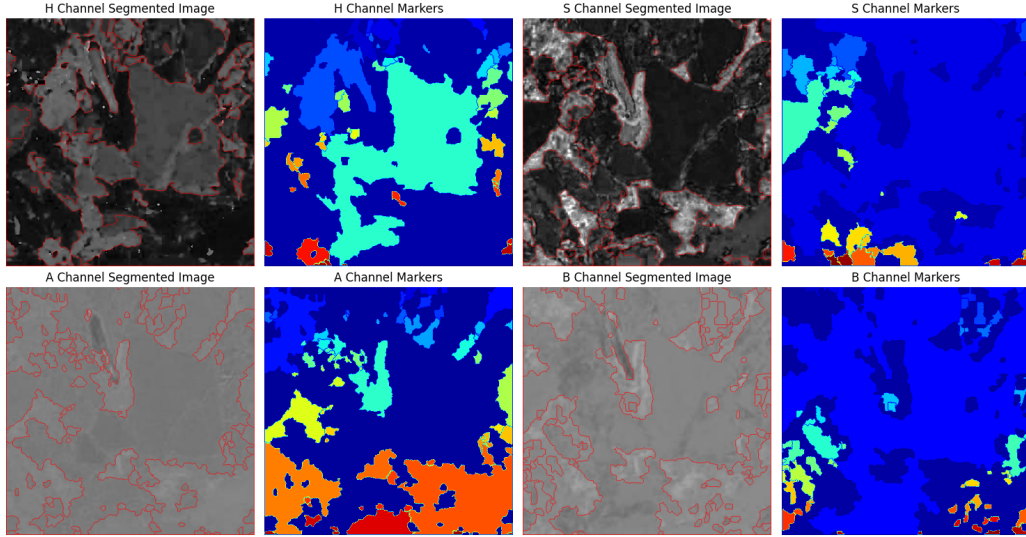


Figure 13: Watershed results applied to the first CPOL image using H, S channels of HSV and A, B channels of LAB colour spaces

6.2 Limitations

- **Lack of labelled data:** The primary limitation of this study is the absence of ground truth data for direct comparison of results. This issue could be addressed by having an expert assess the segmentation outcomes or by improving the available training data with high-resolution grain maps for selected samples. Increasing the availability of training data would be preferable, as it would reduce reliance on human expertise in the data processing pipeline.
- **Limited number of samples:** Due to time constraints within the DSG, only a small number of images could be analysed. Consequently, the findings may not be fully representative of the entire population of sample images.

6.3 Future Work

This investigation has demonstrated that the watershed algorithm can effectively segment grains from thin section scanner images, with its performance being highly sensitive to parameter settings.

- **Improving sure background designation:** The designation of the sure

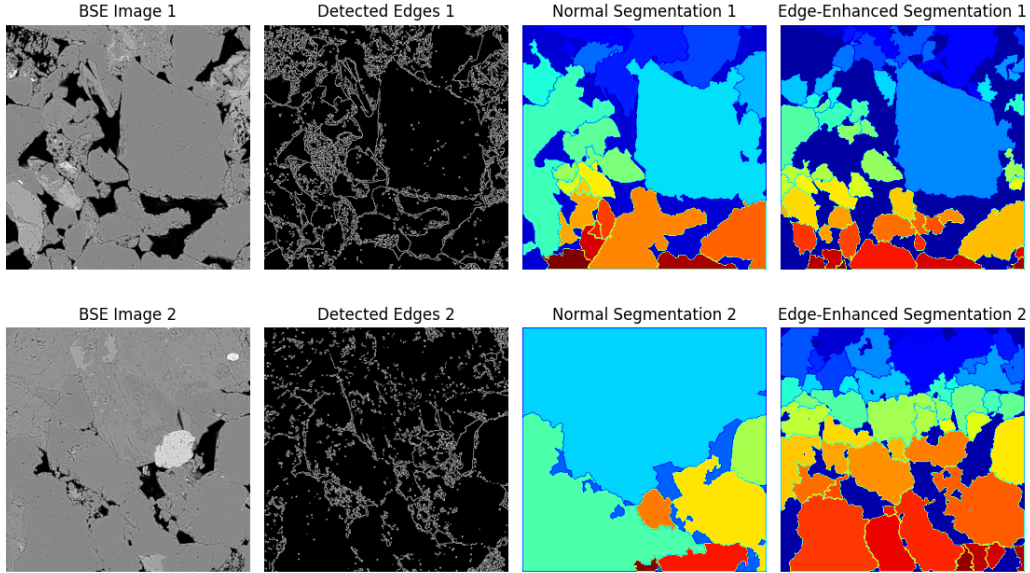


Figure 14: Watershed results for BSE images of the two samples. BSE images (greyscale), detected edges using the Canny algorithm, watershed results without edge enhancement and watershed results using edge enhancement.

background is crucial, and the distinction between pores and grains may be clearer when using a different input image, such as PPL instead of CPOL. Additionally, the sure background designation could be applied to a different image than the one used for the other watershed algorithm steps.

- **Utilisation of multiple image layers:** Various image layers could be leveraged to identify markers for the basin step in the watershed algorithm. For example, differences between XPOL layers could be used to enhance edges, or other combinations of layers could be explored. Incorporating this additional information may help optimise the algorithm's performance, achieving a better balance between under-segmentation and over-segmentation.

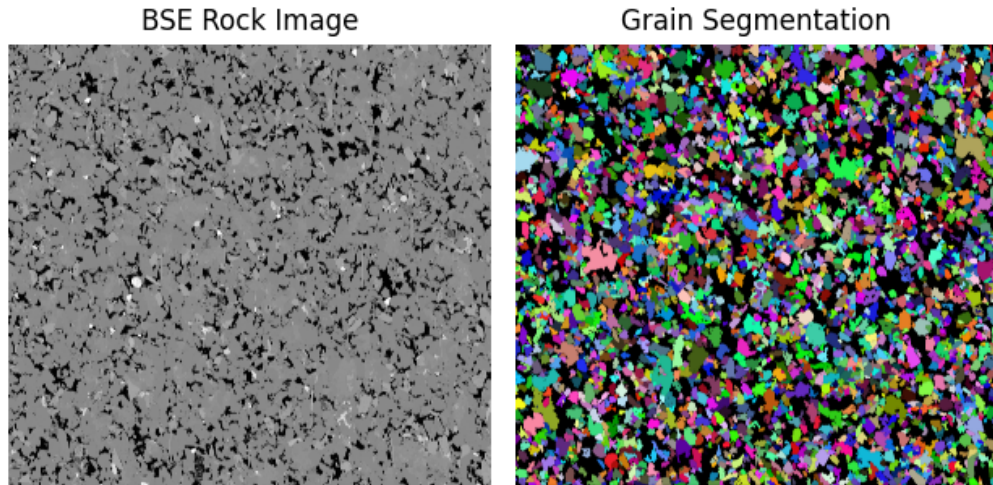


Figure 15: Watershed applied to larger section of the BSE image

7 Mineral Classification

7.1 Random Forest

In this section, we describe the process of constructing ‘signatures’ for different minerals based on their optical properties and training a Random Forest algorithm to classify grains using selected features. The Random Forest algorithm was chosen due to its ability to identify complex relationships in high-dimensional data and its strong performance on similar tasks [25]. The ‘signatures’ consist of statistical summaries that characterise the appearance of different mineral grains under varying light conditions, effectively serving as a feature selection step in constructing the classification model.

The Random Forest algorithm is an ensemble learning method used for classification and regression tasks. It operates by constructing multiple decision

trees during training and, for classification, outputs the mode of the predicted classes. This approach is advantageous in reducing overfitting and handling large datasets with high dimensionality. By aggregating the predictions of multiple trees, Random Forest performs well with non-linear data and captures complex interactions among features, which are internally ranked in terms of importance [6].

The Random Forest algorithm generates multiple bootstrap samples from the training dataset. For each sample, a decision tree is constructed by randomly selecting a subset of features at each node to determine the optimal split. Each tree is grown to its maximum depth without pruning, ensuring that the data is fully utilised. Once the trees are combined to form the forest, classification predictions are made through majority voting [6].

7.1.1 Data Preprocessing

A function was developed to extract three-channel colour data from regions of interest (ROIs) in both Red-Green-Blue (RGB) and Hue-Saturation-Value (HSV) colour spaces to serve as features for the Random Forest model. This was applied to previously categorised ROIs, where all mineral classes were identified using the phase map—a colour-coded classification image.

A data mask was generated based on the colour map to isolate each unique mineral class within the ROI. Three-channel colour data was then extracted from the masked regions under different lighting conditions to selectively collect data for one mineral class at a time. An example of this process is illustrated for the most common mineral class in the samples, quartz, represented as an RGB histogram under six different polarisations of light in Figure 16.

The population data for each mineral class and colour channel was statistically characterised using the mean, median, mode, variance, skewness, and kurtosis for each of the six cross-polarisation angles.

Additionally, for each mineral class, colour data was used to generate channel-specific histograms in both RGB and HSV colour spaces. As a representative example, the mineral class quartz is shown as an HSV histogram under six polarisations of light in Figure 17. This process resulted in three unique characteristic curves per mineral, under six polarisation angles, across two distinct colour spaces. These statistical and histogram-based features collectively serve as the feature space for the Random Forest algorithm.

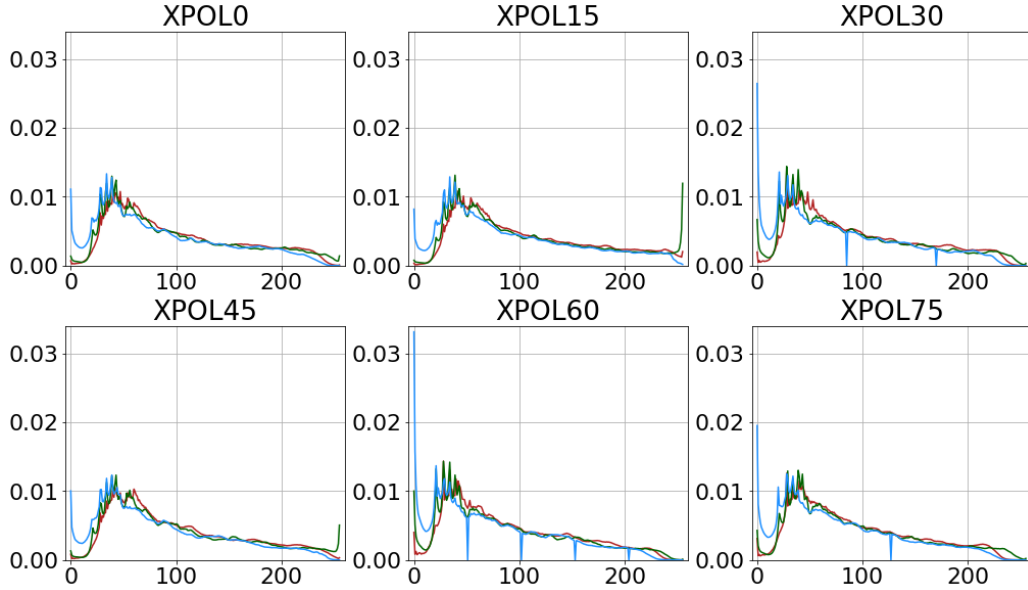


Figure 16: RGB channel histogram for mineral class quartz, as determined by mask-isolated colour channel separation applied to an already characterised thin-slice rock sample. Zero values occurring at regular intervals were observed in some samples, under some polarisations of light. The reason for this may be due to pleochromism (refractive effects in mineral grains) or image processing artefacts.

Each histogram curve exhibited characteristic peaks, with the magnitude and location of the three most prominent peaks identified and recorded. Additionally, the channel values at the minimum and maximum ranges were extracted for each set of channel-specific histograms in both colour spaces.

This process was conducted for all seventeen mineral classes, including porosity, which was treated as a mineral class. However, two mineral classes were excluded from further analysis beyond histogram evaluation due to an insufficient number of pixels available for training. In total, thirteen distinct characterisation features were used for each colour channel and polarisation across all mineral classes.

7.1.2 Model Training

A Random Forest algorithm was trained to identify minerals based on a selection of pixels using the colour features described above. We tested tile sizes of 4096×4096

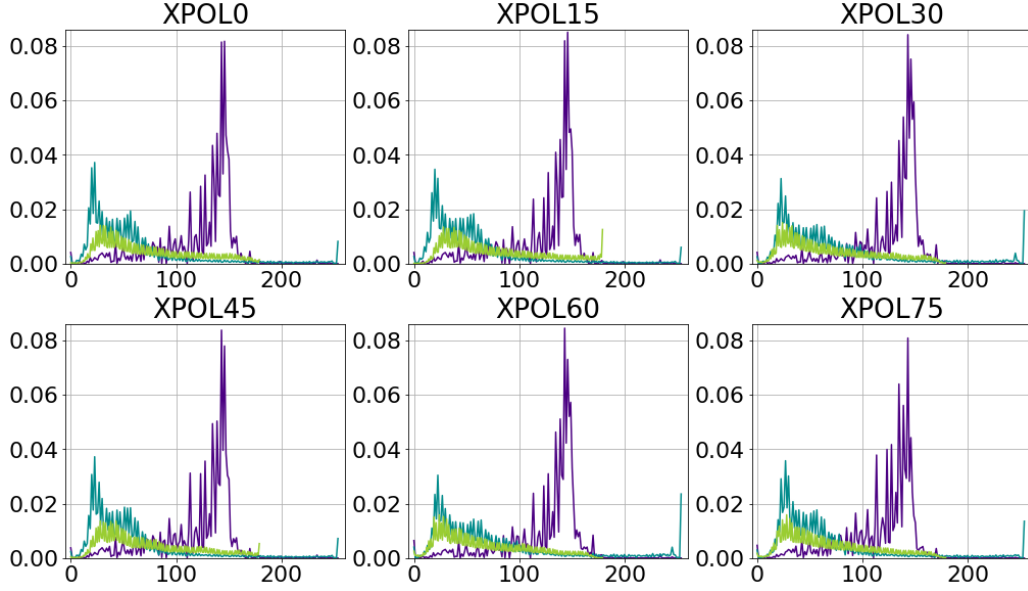


Figure 17: HSV channel histogram for mineral class quartz, (hue is purple, saturation is blue, and value is green), as determined by mask-isolated colour channel separation applied to an already characterised thin-slice rock sample. It should be noted that the value, V channel is represented on its natural scale range of 0 to 180 whereas hue, H and saturation, S have natural scale ranges of 0 to 255.

pixels and larger tiles of 8192×8192 pixels. The set of tiles extracted from the training XPOL images was split into training and testing datasets using an 80:20 ratio. Due to the relatively small number of tiles available from the regions of interest (ROIs), the final training set, after data cleaning, consisted of 236 samples for the 8192 tile size and 1076 samples for the 4096 tile size. The Random Forest models were trained separately in both the RGB and HSV colour spaces.

Due to a processing issue identified after the DSG week, the results for the larger tiles are only available for a test using XPOL75 images, rather than across all image layers.

The algorithm was implemented using SciPy [29]. Given the relatively small training datasets, the number of tree estimators was set to 100 for the smaller training set and 300 for the larger one. Node splitting within each tree was performed using the Gini splitting criterion, with no restrictions on leaf nodes. The maximum number of features considered at each split was set to $\sqrt{N}$, where N

represents the total number of extracted features ($N = 234$ for this Random Forest model, comprising thirteen statistical features for each of the three colour bands across six polarisations of light).

7.1.3 Results

Given the limited number of samples in the raw training set, we applied a sliding window method to tile the ROIs, generating a larger set of training points for the Random Forest algorithm. Unlike convolutional neural network (CNN) input preprocessing, where smaller tiles are often beneficial, the challenge in this case was ensuring that the tiles were sufficiently large for the extracted colour histograms to be representative of the entire ROI. We considered two tile sizes: 4096×4096 pixels and 8192×8192 pixels.

Examples of histograms for a well-represented mineral class and an underrepresented mineral class are presented in Figures 18 and 19, respectively. Even for the well-represented mineral quartz, the histogram changes as the tile size decreases, particularly in the peak locations. For underrepresented classes such as calcite, the summary statistics may vary significantly with smaller tile sizes, making them difficult to identify based solely on histogram features.

Notably, if a representative histogram can be obtained from the training images, converting images into the feature space implicitly balances the classes by discarding information about pixel counts per class. Additional processing of the histograms, such as smoothing over bins, may improve the robustness of the histogram estimates; however, it is essential to preserve the peak locations.

The tiles were then processed following the procedure described in Section 7.1.1, resulting in a set of summary statistics and histogram features for each class, polarisation phase, and tile. These tables were combined into a single feature table to be used as input for training the Random Forest algorithm. By analysing the histograms, we determined that two mineral classes (barite and chlorite) were poorly represented in the raw images, making them unidentifiable based on their histograms. Consequently, these two classes were excluded from the classification, leaving fifteen classes for the final model.

One key advantage of using the Random Forest algorithm in feature space is its ability to rank features by importance, indicating the contribution of each feature to the classification decision. The importance scores for the top 15 features in each trained model are presented in Figure 20 for the RGB colour space and Figure 21

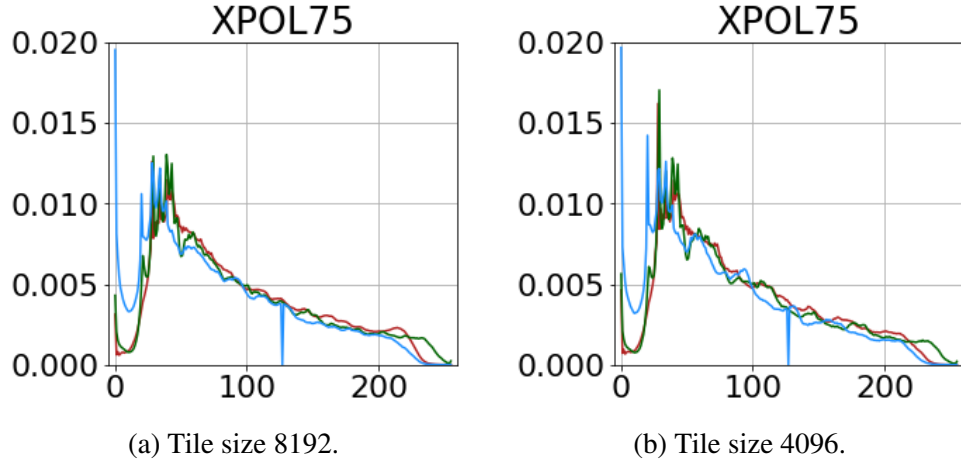


Figure 18: Histograms for Quartz class in (a) tile size 8192, (b) tile size 4096. The histogram is extracted from an image with polarity phase 75 degrees.

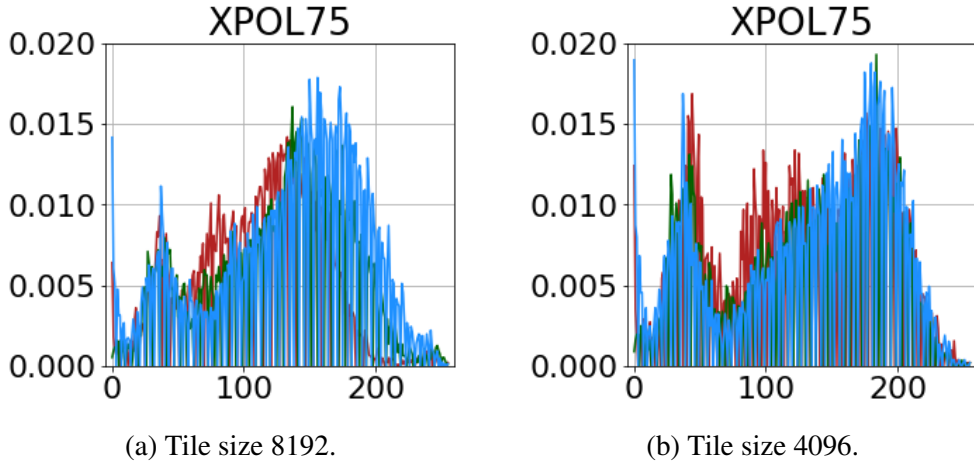


Figure 19: Histograms for Calcite class in (a) tile size 8192, (b) tile size 4096. The histograms were extracted from an image with polarity phase 75 degrees.

for the HSV colour space. Notably, feature importance varies with tile size. In the RGB colour space, the most significant feature for the smaller tile size is the end height (YatEnd), whereas for the larger tile size, variance emerges as the most important feature. In the HSV colour space, the set of highly significant features changes entirely with tile size, with central moments becoming dominant for the larger tiles.

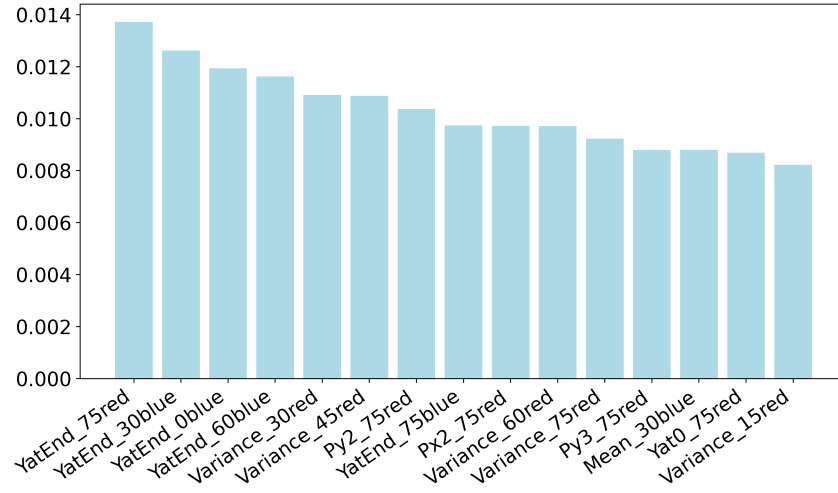
In both colour spaces, the gradual decline in feature importance beyond the top two or three dominant features suggests that the additional features contribute only marginally to information gain. This indicates that, given the current size of the training dataset, reducing the number of features while maintaining classification performance remains challenging. However, if additional training data becomes available, the feature importance rankings could be used to refine the model by reducing its dimensionality while preserving accuracy.

The RF models were trained on datasets corresponding to each of the tile sizes. The confusion matrix obtained from running the trained RF model using HSV and 4096 tile size on test datasets is displayed in Figure 22. There is a strong diagonal of the confusion matrix indicating that many of the minerals are correctly classified using the random forest algorithm. We note that there is an ‘unidentified’ class within the original data, and some of these regions are predicted as other minerals. These ‘false’ predictions may not be incorrect as the original data does not define a class for all of the pixels. Confusion matrices for the other models are located in the Appendix.

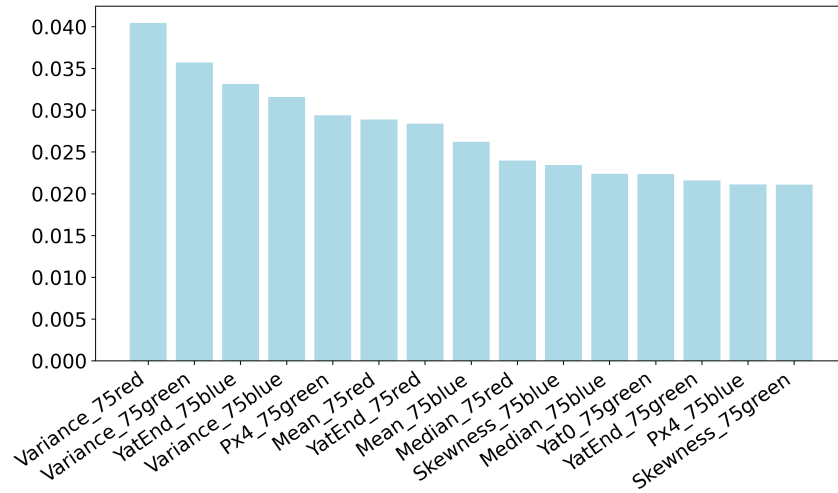
Table 2 presents detailed results for the HSV models. The accuracy scores are 0.74 for the smaller tile size and 0.77 for the larger tile size, demonstrating the effectiveness of the Random Forest model. Certain classes, such as clay matrix, oxide, and unidentified minerals, exhibit relatively low F1 scores compared to other classes. This suggests that these classes have high variability in optical properties across different samples, making them more challenging to classify consistently. Additionally, support is lower for the larger tile size, as fewer test tiles were available. Some mineral classes, such as halite at the 8192×8192 tile size, are underrepresented in the test dataset, highlighting the need for future testing on a more comprehensive dataset. The class imbalance in the test set is accounted for in the weighted average results, which closely align with the macro average results computed across all classes.

7.1.4 Limitations

- Choice of features for random forests: we have distributional information about the colourspace in the form of histograms, so it would be natural to look at Bayesian classifiers that incorporate uncertainty into the classification.
- The selection of attributes for classification using the Random Forest algorithm effectively restricts its application to cases where grain boundary



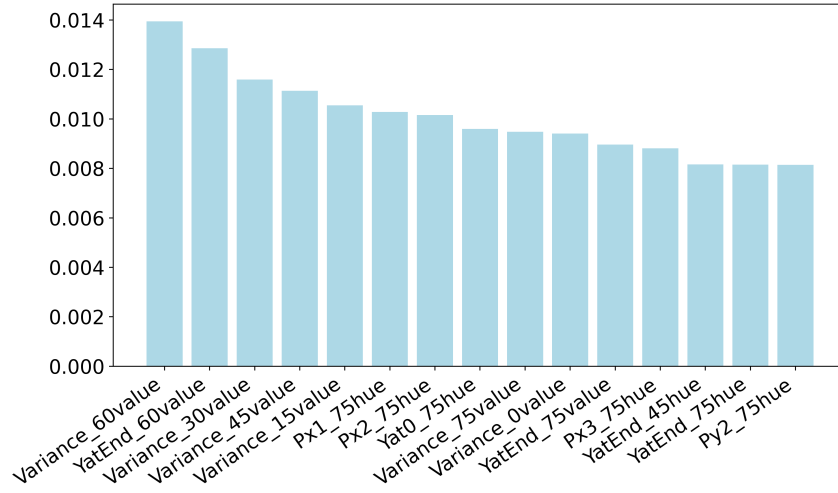
(a) RGB colour space, 4096 tile size.



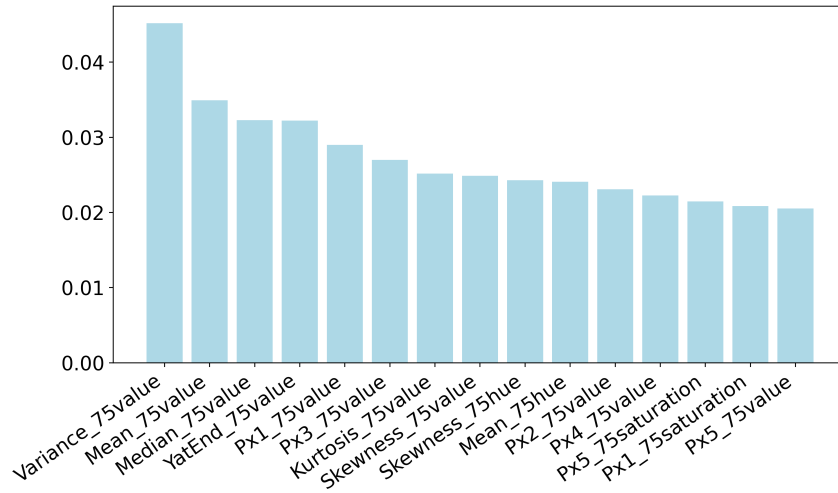
(b) RGB colour space, 8192 tile size.

Figure 20: Feature importance scores for the first 15 features in the random forest models trained on RGB data from datasets obtained from two tiling configurations. Note that only XPOL75 features were used to train the 8192 tile model.

masks are available, as it relies on extracting distributional information from individual grains. Moreover, operating solely on colour features



(a) HSV colour space, 4096 tile size.



(b) HSV colour space, 8192 tile size.

Figure 21: Feature importance scores for the first 15 features in the random forest models trained on HSV data from datasets obtained from two tiling configurations. Note that only XPOL75 features were used to train the 8192 tile model.

prevents the model from incorporating any textural or structural information that may be present in the full image. This suggests that the

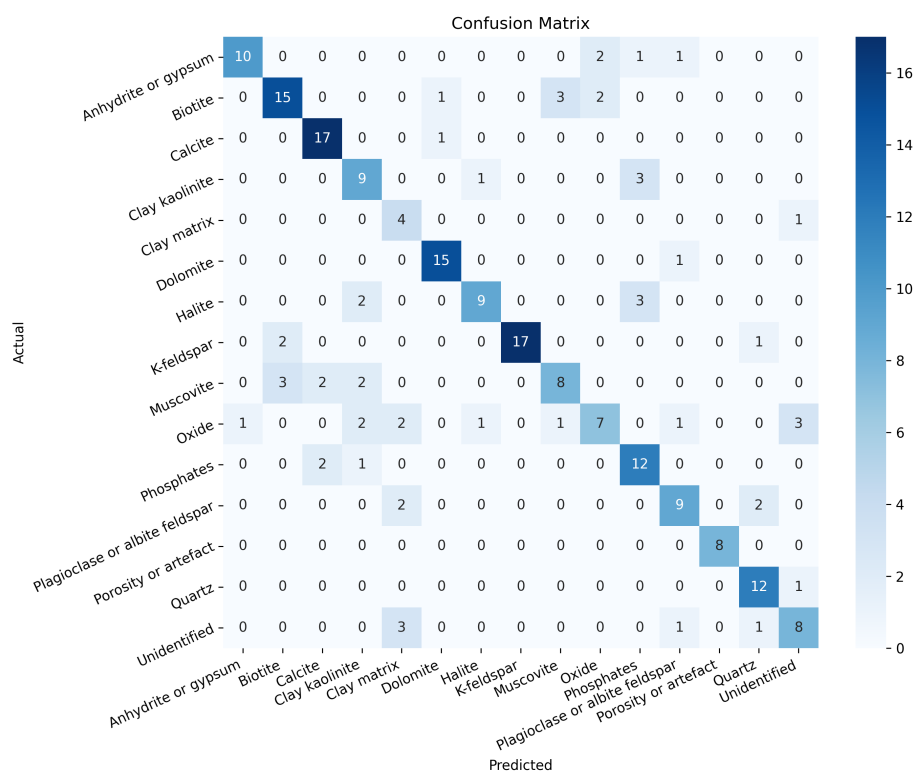


Figure 22: Confusion matrix of the test outputs for the HSV model trained using 4096 tile size.

Table 2: Combined RF classification results (precision, recall, F1 score and support) for the test datasets in HSV colour space at 4096 and 8192 tile sizes.

Class	HSV 4096				HSV 8192			
	Prec.	Recall	F1	Supp.	Prec.	Recall	F1	Supp.
Anhydrite or gypsum	0.89	0.57	0.70	14	1.00	0.75	0.86	4
Biotite	0.75	0.71	0.73	21	1.00	0.75	0.86	4
Calcite	0.81	0.94	0.87	18	1.00	1.00	1.00	4
Clay kaolinite	0.57	0.62	0.59	13	1.00	1.00	1.00	3
Clay matrix	0.36	0.80	0.50	5	0.33	0.50	0.40	4
Dolomite	0.79	0.94	0.86	16	0.80	1.00	0.89	4
Halite	0.67	0.86	0.75	14	1.00	1.00	1.00	1
K-feldspar	1.00	0.90	0.95	20	0.86	1.00	0.92	6
Muscovite	0.70	0.47	0.56	15	1.00	1.00	1.00	5
Oxide	0.54	0.39	0.45	18	0.50	0.50	0.50	2
Phosphates	0.92	0.80	0.86	15	1.00	1.00	1.00	3
Plagioclase or albite feld.	0.73	0.62	0.67	13	0.33	0.40	0.36	5
Porosity or artefact	1.00	1.00	1.00	8	0.67	1.00	0.80	2
Quartz	0.71	0.92	0.80	13	1.00	0.86	0.92	7
Unidentified	0.57	0.62	0.59	13	0.50	0.29	0.36	7
Accuracy	0.74				0.77			
Macro avg	0.73	0.74	0.72	216	0.80	0.80	0.79	61
Weighted avg	0.75	0.74	0.73	216	0.79	0.77	0.77	61

dimensionality reduction process results in significant information loss. Enhancing this classification approach by integrating additional features, such as the geometrical properties of grains, could improve performance.

- Grain segmentation from polarised light images presents significant challenges due to artefacts that arise when observing a three-dimensional object in two dimensions. For instance, variations in texture within a large grain can be misinterpreted as a boundary between separate grains, leading to an overestimation of granularity. Furthermore, the presence of many small grains can negatively impact Random Forest performance in the subsequent processing step, as the colour histogram of a small region tends to be highly ‘spiky’. A post-processing pipeline may be necessary to filter out grains that are too small or too large, incorporating expert knowledge about geometric grain properties to refine the classification process.

7.1.5 Future Work

The Random Forest model for rock property characterisation has demonstrated promising performance, considering the time and resource constraints of the DSG challenge. The model is designed to integrate closely with the grain segmentation module, and its performance is expected to improve once both components are successfully merged. The RF model developed during the DSG challenge provides a highly adaptable framework for sample characterisation. Several avenues warrant further investigation and refinement to enhance model functionality and classification accuracy.

- In the current prototype, the RGB colour space has been the primary source of histogram data from which characterisable features were derived. Late in the development process, HSV colour space was incorporated. Further experiments should be conducted to evaluate the model's performance under different training conditions and with varied feature sets. Although the software is fully functional, its performance has not yet been rigorously validated.
- The inclusion of additional colour spaces has been shown to improve performance, with published literature suggesting that the HSI (hue, saturation, intensity) colour space provides better differentiation between similar minerals [19].
- Some studies have favoured the CIELab colour space over RGB due to its perceptual linearity and device independence, which may be beneficial for characterising samples digitised using different imaging systems [4, 3].
- Histogram colour information from multiple colour spaces could potentially be used simultaneously as features to assess their combined influence on model performance.
- In the current preprocessing routine, peak data from histogram curves is measured as a pure magnitude (from $y = 0$). However, peak values would be more meaningfully ranked if measured relative to the 'curve shoulder,' as is standard in chromatographic analysis. Implementing a peak detection function using a dedicated package such as `hplc-py` [8] instead of `SciPy` would be a straightforward upgrade that could enhance peak selection and ranking accuracy.
- Previous studies have demonstrated high classification success rates using

convolutional filters to extract features from thin-section samples, achieving strong accuracy when classified by neural networks or Random Forest algorithms [25]. This presents an opportunity to integrate methodologies explored in this study group to evaluate their combined effectiveness.

- Ray tracing, if available, could be incorporated into the analysis of BSE images, enabling further analysis using RF. Prior research has shown that pleochroism and interference colours can obscure the ‘true’ colour spectra of reflected light from mineral grains, and ray tracing may help mitigate these effects [2, 1].

Next we present an alternative approach for the task of mineral classification, using the U-Net convolutional neural network model rather than a random forest.

7.2 U-Net

We also approached the challenge of mineral segmentation from a Deep Learning perspective. We chose the U-Net architecture [24] due to its wide applicability across many different segmentation tasks [17]. U-Net is a convolutional neural network originally designed for biomedical image segmentation [24]. It has been shown to be effective in tasks where precise localisation and pixel-wise feature predictions are needed. Previous applications include the identification of boundaries for objects in medical images. U-Net has a symmetric, U-shaped structure, with one contracting path to capture context and a symmetrical expanding path for localisation. The structure enables U-Net learning to incorporate abstract features as well as fine-grained details.

In the contracting path, a series of convolutional layers downsample data to capture abstract representations of the data, which reduces the spatial dimensions and increases depths of feature maps. In the expanding path features are upsampled and correlated with their corresponding layers from the contracting path, where the correlations ensure that spatial information is retained. The output is a segmentation map, where each pixel is assigned a class label.

7.2.1 Data Preprocessing

The initial exploration of the training images revealed that the phase maps (which comprise the ground-truth labels for the data) were available only for subsections

of each sample (Figure 2(b)). The phase maps had variable locations, shapes, and orientations within each sample. The initial pre-processing step therefore involved locating the phase map within the sample as a region of interest (ROI), and identifying them with bounding boxes. Smaller patches (256×256 pixels, for example) could then be extracted from these ROIs for further analysis. This step informed which areas were used for all four image-processing work packages.

To train the U-Net algorithm, the images from an ROI were combined into a single array by stacking the different images into a single tensor along the colour channels. Then the stacked arrays were split along the shape dimensions into a set of 256×256 non-overlapping images. This process yielded 13272 training samples from 7 thin sections, 1680 test samples from three bounding boxes of a single section, and 438 validation samples from a single section. The samples were normalised along the colour channel dimensions. The input tiles that are selected for training undergo an augmentation process: randomised linear transforms of the image shape dimensions (including flipping, rotating, cropping) were applied to each training tile with a probability 0.5.

7.2.2 Model Training

Our training pipeline was implemented in Python using the PyTorch [22] and Segmentation Models² libraries. It is a common practice in computer vision to configure the U-Net architecture to follow an established model architecture that has demonstrated strong performance in related tasks. Following this approach, we selected the encoder of our model to follow the ResNet34 architecture [12], with model weights initialized using values pre-trained on the ImageNet dataset.

Prior to training, the data was split into 256×256 pixel square tiles, and 18 masks were extracted, corresponding to the 18 classes represented by different colours in the phase map instances of the dataset. Notably, there is one additional class compared to the RF processing, as ‘white space’ was included to indicate areas without phase map information.

During training, images from three different light conditions (CPL, XPL0, and PPL) were loaded and stacked to form a tensor of shape $B \times C \times H \times W$, where B represents the batch size, C denotes the number of colour channels (nine, derived from three RGB images), and H, W correspond to the height and width dimensions

²Segmentation Models Git Repository

of the input images. The labels consist of the 18 masks derived from the phase map images, where colours were converted into class labels, forming 18 binary arrays that indicate the presence or absence of each mineral type.

The model was trained for 40 epochs with a batch size of 16 data instances. We employed the Adam optimiser [16] to update the network parameters at each epoch. For the loss function, we used binary cross-entropy loss (BCE). BCE is a standard choice for binary classification tasks, where the objective is to assign inputs to one of two categories. Given our dataset format (18 binary masks), BCE is an appropriate optimisation criterion. The BCE loss measures the difference between the true labels and the predicted probabilities, providing a quantification of how well the model’s predictions align with the actual labels.

It is given by the following formula:

$$\text{BCE Loss} = -\frac{1}{N} \sum_{i=1}^N [y_i \cdot \log(p_i) + (1 - y_i) \cdot \log(1 - p_i)] \quad (1)$$

where N is the batch size in pixels, y_i is the true label (binary for each class) for pixel i , and p_i is the prediction output for pixel i . This loss function is averaged over the 18 classes. Training was evaluated on the fly with a validation set from an image outside the training set to prevent data leakage.

7.2.3 Results

For the mineral segmentation task, we chose to evaluate our method with the F1 Score. The F1 score is a metric commonly used to evaluate the performance of classification models, particularly when dealing with imbalanced datasets. It is the harmonic mean of Precision (proportion of correct classifications compared to the total classifications for each label) and Recall (proportion of correct classifications compared to true label size), providing a single metric that balances both concerns. The F1 Score ranges from 0 to 1, where 1 indicates perfect precision and recall, and 0 indicates the worst possible performance.

The F1 Score is calculated as:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (2)$$

In our case since we are dealing with 18 classes, the random prediction (assuming equally sized classes) would yield an F1 score of 5.834%. We report an F1 score of

27.992% for our U-Net, which is 5-fold improvement despite training for a small number of epochs.

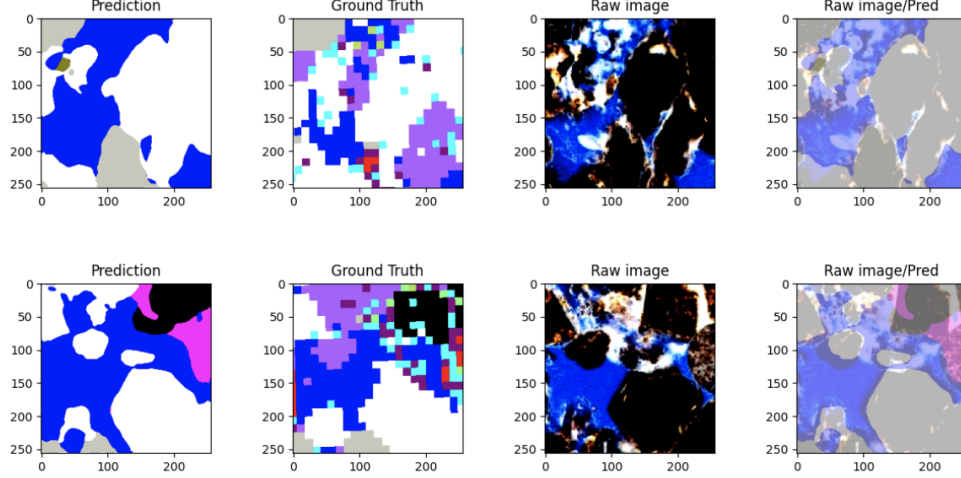


Figure 23: Example outputs from the U-Net model, consisting of the predicted phase map, ground-truth phase map, raw image input, and a super-position of the predicted phase map with the raw image input. Note that the colours used to denote classes here do not match the original phasemap colours.

We also report the Intersection over Union (IoU) score, also known as the Jaccard Index, which is a metric used to evaluate the performance of segmentation models by comparing the overlap between the predicted segmentation mask and the ground truth mask. IoU measures the ratio of the area of overlap between the predicted and ground truth masks to the area of their union. The IoU value ranges from 0 to 1, where 1 indicates perfect overlap (i.e., the predicted and ground truth masks are identical), and 0 indicates no overlap. The IoU is calculated as:

$$\text{IoU} = \frac{\sum (P \cdot G)}{\sum P + \sum G - \sum (P \cdot G)} \quad (3)$$

where P is predicted and G the ground truth. We report an IoU score of 29.865%.

Two example outputs are shown in Figure 23. The ground truth images demonstrate the highly pixelated nature of the labelled training data, however the predictions show smooth edges. This shows that the network is not overfitting to create blocky

outputs to match the inputs. The prediction regions align well with the shapes on the raw image for a first training attempt, and further training and tuning of hyperparameters should improve this model.

The confusion matrix for the test set is shown in Figure 24. The log scale for the colour was chosen as the most abundant classes were completely dominating the image in linear colour scale leaving most of the image pale. The U-Net looks to predict a subset of the classes, with darker pixels on the diagonal correct classifications. One factor to alter in future training is the handling of class imbalance, so that the network predicts each of the relevant classes. As in the RF classification the unidentified class does not give a representative indication of the performance of the network, and the high valued pixels on the unidentified prediction column could indicate that some originally unidentified classed pixels appear the same as phosphate, porosity and quartz. As the RF and U-Net algorithms were trained and tested using different tile sizes it is not meaningful to directly compare the classification metrics. The U-Net algorithm uses only the images as input, whereas the RF requires knowledge of the grain boundaries as a classification is evaluated for each grain.

7.2.4 Limitations

Data limitations:

- The size of the raw images had presented the team with a significant challenge as it rendered a lot of traditional data exploratory approaches inapplicable. Cropping of original files into manageable slices was not straightforward because the borders of the phasemap ROIs were misaligned with the image axes. As a result, the information about underrepresented classes in mineral classification task had become available when the U-Net framework was already set up for the full set of classes.
- The phase maps that contained label information about mineral classes had different resolution from the polarity images, which resulted in the class masks being rather ‘blocky’. We did not address this resolution difference during the DSG, and treated the phasemaps as if they had the same resolution as the images. Both classification models explored during the study group alleviate this problem. U-Net incorporates object segmentation and has returned more precise object borders than the training data phasemap even after the short training period. The feature-based classification deals with

distributional information and thus the colour values inaccurately classified pixels constitute only a small subset of the sample used to build a histogram (these values are likely to end up in the tails of colour distributions).

- Some identified minerals were not present in the provided samples, which could therefore not be used for training the models. In practice, training should include samples that contain minerals of concern to facilitate their identification in test samples.

Approach limitations:

- The fundamental limitation of any deep learning algorithm is the lack of interpretability. While the output of the framework returned smoother borders of grains than the available labelling image, without adding explainable AI methods there is no way to determine what information contained in any given pixel contributes to the decision about its class. In the future, the incorporation of an attention mechanism may improve user understanding.

Technical limitations:

- **Resource Constraints:** The provided virtual machines had a small number of CPUs that led to dataloader issues while training. Namely, a 64 batch size training instance required 178 minutes to finish a single epoch of training in the provided machines, due to the small number of CPUs available which led to a small number of workers available to handle the data loader fetching.

7.2.5 Future work

The preliminary analysis we present shows promising results for future work. We managed to classify well-represented classes that are prominent in the phasemap images successfully, however class imbalance affected the accuracy of the model in under-represented classes. Several approaches could be taken to improve the performance of grain segmentation with the U-Net model:

- **Class imbalance:** There is a substantial performance fall in prediction accuracy for under-represented classes. A way to deal with this issue is data augmentation focused on oversampling the minority classes via techniques (e.g. rotations, flips, scaling) specifically applied to the samples of those classes. This generates more diverse examples, helping the model learn better features for minority classes. Weights can also be applied to the

loss function within the network so that under-represented classes are up-weighted in the loss function.

- **Training on overlapping images:** Our dataset is generated by cropping square tiles from the initial data. Training a segmentation model on overlapping images, as opposed to non-overlapping images, can offer several benefits, particularly in terms of model performance and generalisation. Firstly, creating a dataset from overlapping images would generate more training data and preservation of contextual information that may be lost in the non-overlapping case. Additionally, overlapping would assist in mitigating boundary artefacts, which occur when the model makes errors at the edges of non-overlapping tiles due to the lack of surrounding context.
- **Time/resource constraints:** Resource-wise our progress was hindered by the bottleneck caused by the CPU-to-GPU processing pipeline, which significantly slowed down the training process. This bottleneck limited our ability to fully utilise the available GPU resources, leading to extended training times. As a result, we were unable to explore different model configurations and conduct an extensive parameter search within the project timeline. Availability of more resources will enable to experiment with a broader range of model architectures and hyperparameters, ultimately improving the performance and generalisability of the model.

8 Discussion

Within this challenge, we investigated a range of methods for the different tasks, from simple rule-based approaches to neural network models. The variety of techniques explored highlights the potential for combining methods from different disciplines to construct a single data processing pipeline.

For the grain classification problem, two methods were investigated: a feature-based random forest and a U-Net. The feature-based approach closely resembles how a geologist would analyse thin-section images, examining grains and their properties to classify them. However, this method is dependent on grain boundaries being accurately defined beforehand. Despite this limitation, the explainability provided by the random forest approach is superior to that of a neural network model, which processes the image stack as input and produces a

per-pixel classification as output. In contrast, the neural network approach may identify differences between minerals that are not typically recognised or utilised by geologists, as these distinctions are derived from complex combinations of input information.

During the DSG, the two methods were not applied to the same subsets of the dataset, preventing direct result comparisons. Future development should ensure that the same data and training splits are used for each method to enable a direct comparison of their performance.

With a well-trained U-Net, both grain types and porosity could be estimated simultaneously. However, grain boundaries would still require a separate processing step, as adjacent grains of the same type would not be distinguished in a pixel-wise classification. Comparing the computationally inexpensive rule-based porosity estimate with the more complex U-Net estimates across the full dataset would provide insight into the consistency of results across methods.

Due to time constraints, we did not have the opportunity to fully explore the impact of multiple lighting conditions within the methods. For example, the U-Net used three lighting conditions as input, but with greater computational resources and longer run times, all eight conditions could be incorporated.

Features such as grain size, shape, orientation, and texture were not investigated in this DSG. These features depend on accurate grain segmentation, and without a labelled dataset of grain boundaries, this analysis could not be conducted before completing the work on the watershed algorithm.

The performance of the different classification methods varies depending on the approach taken. A key comparison is between making decisions at the grain level versus the pixel level. Below, we discuss key observations regarding the different methods, their relative performance, and the implications of the results.

- **Grain-based vs. pixel-based classification:** The two approaches reflect different conceptual frameworks—the “geologist’s approach” versus the “machine learning approach.” By performing grain segmentation followed by classification, we replicate the classical methodology, which leverages an understanding of the physical properties of rock slices. In contrast, a convolutional neural network (CNN) integrates segmentation and classification into a single step, learning directly from image data. However, if new insights arise in a format other than an image,

incorporating them into a feature-based model is often more straightforward than adapting a CNN.

- **Histogram dips:** Some colour channel histograms for multiple mineral classes exhibit dips to zero at seemingly regular intervals along the channel value range (e.g., the blue channel in the Quartz RGB histograms shown in Figure 16). These dips cannot be attributed to small sample sizes, as the pixel count for the image regions used to generate the histograms can reach the order of 10^6 . This raises the question of whether these dips contain information related to the texture or reflectivity of minerals. Further analysis is required to determine whether these dips occur consistently at the same channel values across different samples for a given mineral class. If they are found to be consistent, they should be included as additional features for training the Random Forest classifiers.

9 Conclusion

Within this DSG, significant progress was made towards developing an automated process for analysing thin-section images and evaluating sites' suitability for carbon or hydrogen storage. The algorithmic steps have been prototyped and are ready for further development by BGS to establish a production pipeline based on this work.

Porosity was estimated using a light blue filter applied to plane-polarised light images, providing an estimate of porosity for each sample. The porosity mask generated in this step can also serve as an estimate of the 'background' area for grain segmentation. Grain segmentation using the watershed algorithm proved particularly effective for larger grains. This segmentation could enhance the phasemap images (which are currently available at a lower resolution than the thin-section images) by providing information about grain boundaries that lie between the lower-resolution pixels. An improved phasemap could be utilised as input for both of the mineral classification algorithms tested during the DSG.

Feature-based random forest classifiers were trained to classify minerals within the images, leveraging the high-resolution original images to identify the visual signatures of each mineral. An alternative approach for mineral classification was also developed using a U-Net algorithm, which utilised visual information from key lighting conditions to generate a pixel-wise classification of the images into

mineral classes.

Further research could enhance the performance of the algorithms prototyped within this DSG, and additional insights could be obtained from this thin-section dataset, particularly by incorporating information from all available lighting conditions. Suggestions for algorithmic improvements and further work are outlined in the methods sections. This Data Study Group has provided important proof-of-concept results within a short timescale, forming a foundation for future research and development.

10 Team members

Samuel Baker is a researcher at the University of Oxford, specialising in machine learning and data processing for turbulent flows. His contributions to the project helped to establish the deep learning pipeline for the mineral identification supervised segmentation task.

Nikos Chazaridis is a PhD student with MINDS CDT at the University of Southampton. His research focuses on deep learning methods to uncover patterns of interest from complex temporal signals with applications in audio and speech. He contributed to this project by implementing the data generation pipeline.

Daniel Dalton is an engineering PhD student at the FIBE2 CDT at the University of Cambridge. His research focuses on optimisation methods of steel structures. He contributed to this project by using colour mapping of images to find porosity of samples.

Lora Frayling is a Data Scientist at Health Data Insight, with an MSc in Data Science from the London School of Economics (LSE). She specialises in synthetic data generation and applying data science methods to healthcare data. In this project, Lora contributed to data preparation for implementing deep learning and machine learning models. She also applied unsupervised clustering algorithms, such as the Watershed algorithm, to segment images of individual grains, successfully identifying the most prominent grain structures. This work can support further analysis of grains and porosity and aid in feature engineering for ML models classifying mineral types.

Anastasia Kadochnikova is a postdoctoral research fellow at the University of

Nottingham. She works at the interface of phenomenological and data-driven methods of modelling dynamical systems. Her research interests include interpretable machine learning and system identification. She implemented the feature-based mineral classification and contributed to the discussion of the report.

Eric Martin is a Lecturer of Civil Environmental Engineering at University of Greenwich with a background in non-isothermal multiphase transport in porous media. He was the Facilitator of this project and contributed to theory, writing and code writing.

Dimitrios Tsiakmakis is a PhD candidate at the BioRobotics Institute of the Sant'Anna School of Advanced Studies in Pisa, Italy. His research lies at the intersection of machine learning and robotics, specialising in capacitive-based sensors applications for proximity tasks in human-robot collaboration. His contributions to the project helped to establish the deep learning pipeline for the mineral identification supervised segmentation task.

Kathryn Leeming is a data scientist at the British Geological Survey with a background in applying numerical approaches to environmental data. She was the PI for this challenge and formulated the geological questions of interest into data questions.

References

- [1] ACEVEDO ZAMORA, M. *Petrographic Microscopy of Geological Textural Patterns and Element-Mineral Associations with Novel Image Analysis Methods*. PhD thesis, Queensland University of Technology, 2024.
- [2] ACEVEDO ZAMORA, M., AND KAMBER, B. Petrographic microscopy with ray tracing and segmentation from multi-angle polarisation whole-slide images. *Minerals* 13, 2 (2023), 156.
- [3] ALIGHOLI, S., KHAJAVI, R., KHANDELWAL, M., AND ARMAGHANI, D. Mineral texture identification using local binary patterns equipped with a classification and recognition updating system (CARUS). *Sustainability* 14, 18 (2022), 11291.
- [4] ALIGHOLI, S., LASHKARIPOUR, G., KHAJAVI, R., AND RAZMARA, M. Automatic mineral identification using color tracking. *Pattern Recognition* 65 (2017), 164–174.
- [5] BRADSKI, G. The OpenCV Library. *Dr. Dobb's Journal of Software Tools* (2000).
- [6] BREIMAN, L. Random forests. *Machine Learning* 45, 1 (2001), 5–32.
- [7] CANNY, J. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 8, 6 (1986), 679–698.
- [8] CHURE, G., AND CREMER, J. hplc-py: A python utility for rapid quantification of complex chemical chromatograms. *Journal of Open Source Software* 9, 94 (2024).
- [9] DANIELS, S., HARDIMAN, L., HARTGILL, D., HUNN, V., JONES, R., AND ROBERTSON, N. Deep geological storage of CO₂ on the UK continental shelf-containment certainty. Supplementary note A: Breakdown of combined well and geological storage risks for typical storage sites. Tech. rep., UK-GOV, OGL, 2023.
- [10] FISHER, A. Cloud and cloud-shadow detection in SPOT5 HRG imagery with automated morphological feature extraction. *Remote Sensing* 6, 1 (2014).
- [11] GREAT BRITAIN. DEPARTMENT FOR BUSINESS, ENERGY AND INDUSTRIAL STRATEGY. Netzero strategy: Build back greener, 2021.

- [12] HE, K., ZHANG, X., REN, S., AND SUN, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 770–778.
- [13] HOLLOWAY, S., ROCHELLE, C., BATEMAN, K., PEARCE, J., BAILY, H., AND METCALFE, R. The underground disposal of carbon dioxide. Tech. rep., British Geological Survey, 1996.
- [14] HOLLOWAY, S., AND SAVAGE, D. The potential for aquifer disposal of carbon dioxide in the UK. *Energy Conversion and Management* 34, 9-11 (1993), 925–932.
- [15] IPCC. Climate change 2021: The physical science basis. contribution of working group i to the sixth assessment report of the intergovernmental panel on climate change. Tech. rep., Cambridge University Press, 2021.
- [16] KINGMA, D. P. Adam: A method for stochastic optimization. *International Conference on Learning Representations 2014* (2014).
- [17] MINAEI, S., BOYKOV, Y., PORIKLI, F., PLAZA, A., KEHTARNAVAZ, N., AND TERZOPOULOS, D. Image segmentation using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 7 (2022), 3523–3542.
- [18] NASA. Climate change: How do we know? <https://climate.nasa.gov/evidence/> (Accessed: 11 September 2024).
- [19] NASERI, A., AND REZAEI NASAB, A. Automatic identification of minerals in thin sections using. *Journal of Ambient Intelligence and Humanized Computing* 14, 4 (2023), 3369–3381.
- [20] ORESKES, N. The scientific consensus on climate change. *Science* 306, 5702 (2004), 1686.
- [21] OTSU, N. A threshold selection method from gray-level histograms. *IEEE Transactions on System, Man, and Cybernetics* 9, 1 (1979), 62–66.
- [22] PASZKE, A., GROSS, S., MASSA, F., LERER, A., BRADBURY, J., CHANAN, G., KILLEEN, T., LIN, Z., GIMELSHEIN, N., ANTIGA, L., DESMAISON, A., KÖPF, A., YANG, E., DeVITO, Z., RAISON, M., TEJANI, A., CHILAMKURTHY, S., STEINER, B., FANG, L., BAI, J., AND CHINTALA, S. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems* 32 (*NeurIPS 2019*) (2019).

- [23] PAYTON, R. L., CHIARELLA, D., AND KINGDON, A. The influence of grain shape and size on the relationship between porosity and permeability in sandstone: a digital approach. *Scientific Reports* 12, 7531 (2022).
- [24] RONNEBERGER, O., FISCHER, P., AND BROX, T. U-Net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention (MICCAI)* (2015), 234–241.
- [25] RUBO, R., DE CARVALHO CARNEIRO, C., MICHELON, M., AND DOS SANTOS GIORIA, R. Digital petrography: Mineralogy and porosity identification using machine learning algorithms in petrographic thin section images. *Journal of Petroleum Science and Engineering* 183 (2019), 106382.
- [26] RUSHTON, J., HANNIS, S., AND MILODOWSKI, A. Petrography and diagenesis of the Bunter Sandstone Formation in the UK Southern North Sea. Tech. Rep. OR/23/054, British Geological Survey, 2023.
- [27] SURVEY, B. G. Geological carbon storage. <https://www.bgs.ac.uk/discovering-geology/climate-change/carbon-capture-and-storage/> (Accessed: 11 September 2024).
- [28] TANG, K. Climate change policies of the four largest global emitters of greenhouse gases: their similarities, differences and way forward. *Journal of Energy Research and Reviews* 10, 2 (2022), 19–35.
- [29] VIRTANEN, P., GOMMERS, R., OLIPHANT, T. E., HABERLAND, M., REDDY, T., COURNAPEAU, D., BUROVSKI, E., PETERSON, P., WECKESSER, W., BRIGHT, J., VAN DER WALT, S. J., BRETT, M., WILSON, J., MILLMAN, K. J., MAYOROV, N., NELSON, A. R. J., JONES, E., KERN, R., LARSON, E., CAREY, C. J., POLAT, İ., FENG, Y., MOORE, E. W., VANDERPLAS, J., LAXALDE, D., PERKTOLD, J., CIMRMAN, R., HENRIKSEN, I., QUINTERO, E. A., HARRIS, C. R., ARCHIBALD, A. M., RIBEIRO, A. H., PEDREGOSA, F., VAN MULBREGT, P., AND SciPy 1.0 CONTRIBUTORS. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* 17 (2020), 261–272.

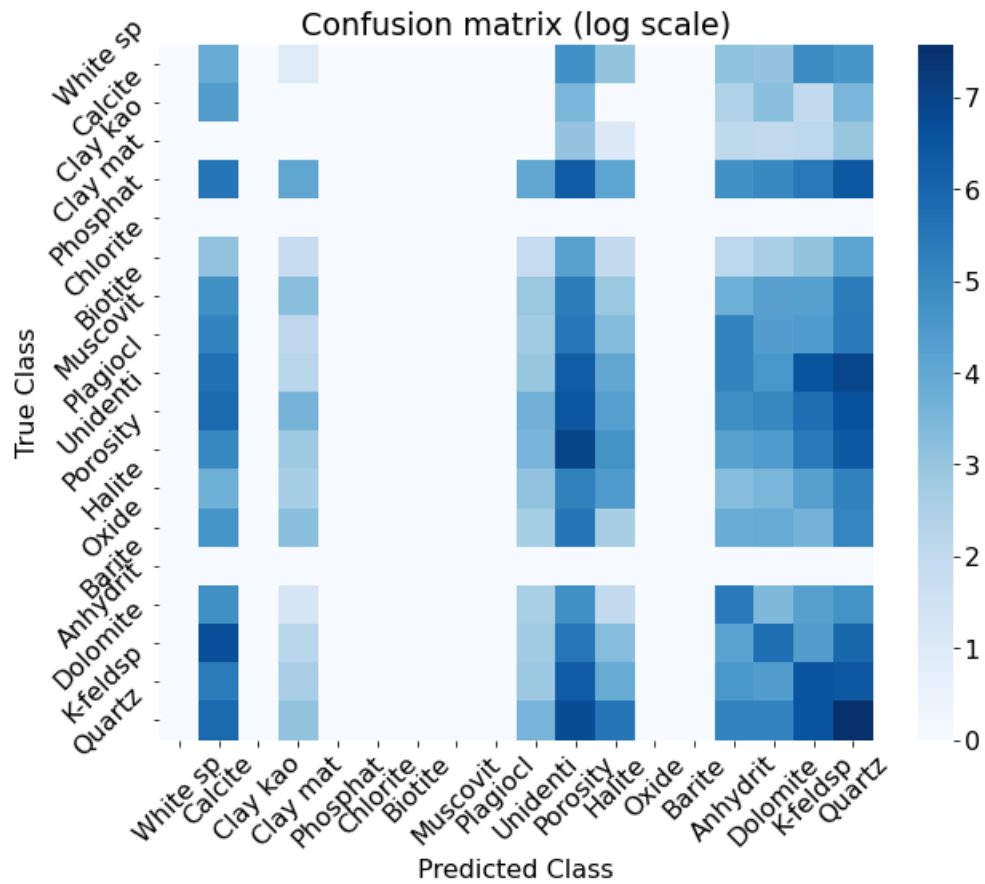


Figure 24: Confusion matrix for the U-Net predictions where the colour map refers to a logarithmic scale, showing a clear class imbalance that should be addressed in future work.

Appendix

Additional figures and tables are included in this section that are supplementary to the main text.

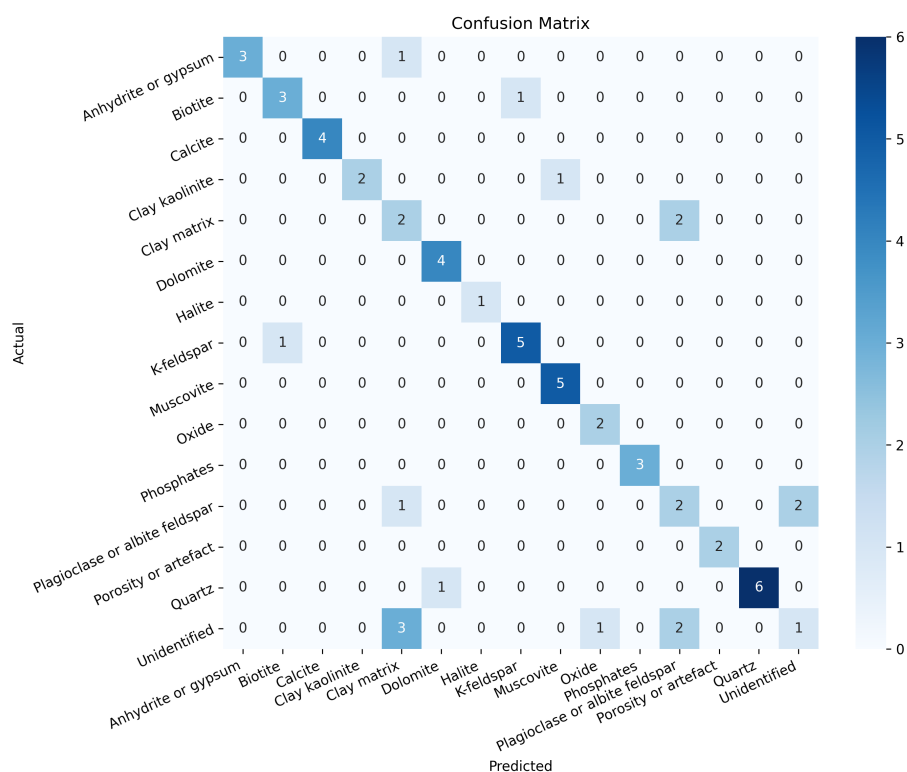


Figure 25: Confusion matrix of the test outputs for the random forest HSV model trained using 8192 tile size.

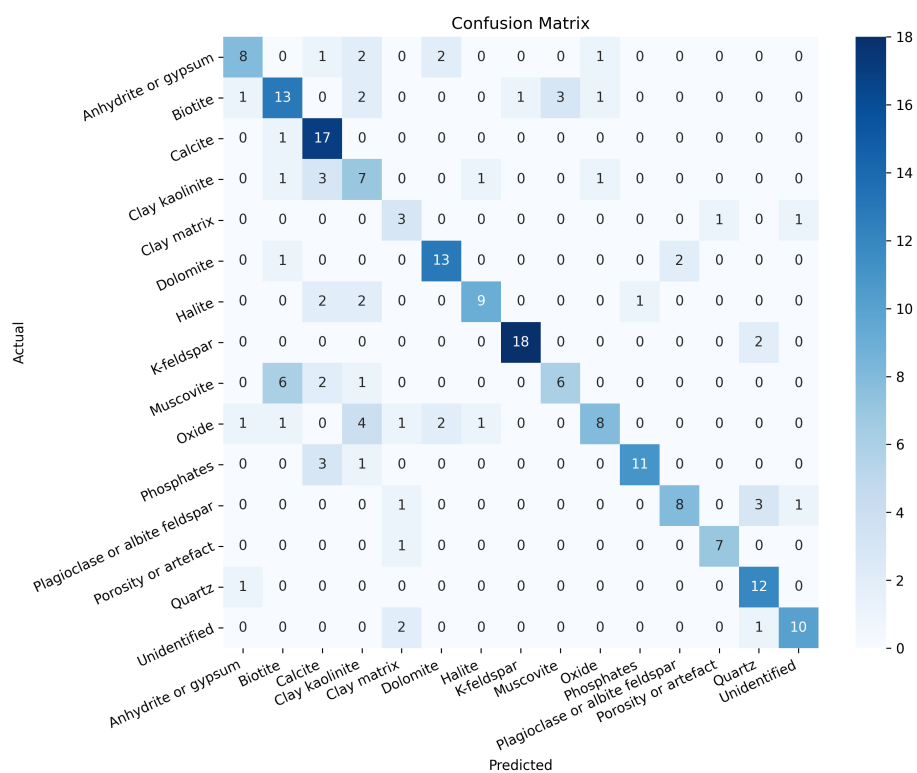


Figure 26: Confusion matrix of the test outputs for the random forest RGB model trained using 4096 tile size.

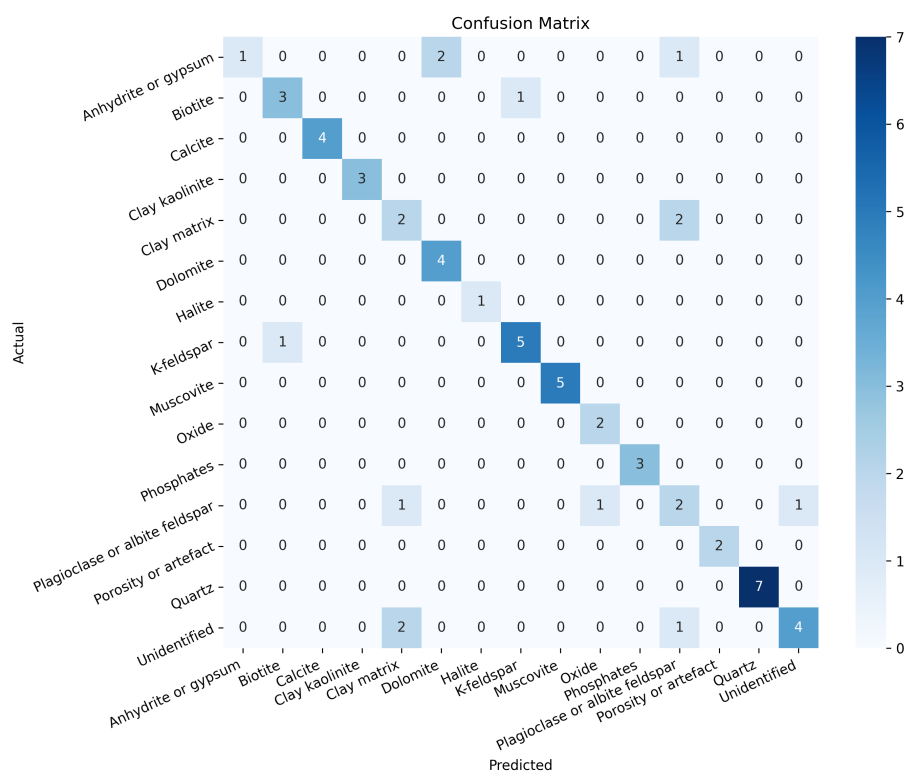


Figure 27: Confusion matrix of the test outputs for the random forest RGB model trained using 8192 tile size.

Table 3: Combined RF classification results (precision, recall, F1 score and support) for the test datasets in HSV and RGB colour spaces at 4096 and 8192 tile sizes.

Class	RGB 4096				RGB 8192			
	Prec.	Recall	F1	Supp.	Prec.	Recall	F1	Supp.
Anhydrite or gypsum	0.78	0.50	0.61	14	1.00	0.25	0.40	4
Biotite	0.63	0.57	0.60	21	0.75	0.75	0.75	4
Calcite	0.61	0.94	0.74	18	1.00	1.00	1.00	4
Clay kaolinite	0.37	0.54	0.44	13	1.00	1.00	1.00	3
Clay matrix	0.22	0.40	0.29	5	0.50	0.50	0.50	4
Dolomite	0.82	0.88	0.85	16	0.67	1.00	0.80	4
Halite	0.82	0.64	0.72	14	1.00	1.00	1.00	1
K-feldspar	0.95	0.90	0.92	20	0.83	0.83	0.83	6
Muscovite	0.58	0.47	0.52	15	1.00	1.00	1.00	5
Oxide	0.62	0.44	0.52	18	0.67	1.00	0.80	2
Phosphates	0.92	0.73	0.81	15	1.00	1.00	1.00	3
Plagioclase or albite feld.	0.73	0.62	0.67	13	0.33	0.40	0.36	5
Porosity or artefact	0.86	0.75	0.80	8	1.00	1.00	1.00	2
Quartz	0.68	1.00	0.81	13	1.00	1.00	1.00	7
Unidentified	0.82	0.69	0.75	13	0.83	0.71	0.77	7
Accuracy	0.69				0.80			
Macro avg	0.69	0.67	0.67	216	0.84	0.83	0.81	61
Weighted avg	0.71	0.69	0.69	216	0.83	0.80	0.80	61



**The
Alan Turing
Institute**

**turing.ac.uk
@turinginst**