https://doi.org/10.5194/egusphere-2023-1819
Preprint. Discussion started: 14 August 2023

Using Internal Standards in Time-resolved X-ray Micro-computed Tomography to Quantify Grain-scale Developments in Solid State Mineral Reactions

Roberto Emanuele Rizzo[1,2], Damien Freitas[2,3], James Gilgannon[2], Sohan Seth[4], Ian B Butler[2], Gina McGill[2,5], and Florian Fusseis[2]

[1]Department of Earth Sciences, University of Florence, Via La Pira 4, 50121, Florence, IT
[2]School of Geosciences, The University of Edinburgh, The King's Buildings, James Hutton Road, Edinburgh EH9 3FE, UK
[3]University of Manchester, Diamond Light Source, Harwell Campus, Didcot OX11 0DE, UK
[4]Data Science Unit, School of Informatics, The University of Edinburgh, 10 Crichton Street, Edinburgh, EH8 9AB
[5]Earth Sciences Institute of Orléans, University of Orleans, T1A Rue de la Férollerie – CS 20066, F-45071 Orléans Cedex 2, France

Correspondence: Roberto Emanuele Rizzo (robertoemanuele.rizzo@unifi)

Abstract. X-ray computed tomography has established itself as a crucial tool in the analysis of rock materials, providing the ability to visualise intricate 3D microstructures and capture quantitative information about internal phenomena such as structural damage, mineral reactions, and fluid-rock interactions. The efficacy of this tool, however, depends significantly on the precision of image segmentation, a process that has seen varied results across different methodologies, ranging from simple histogram thresholding to more complex machine learning and deep learning strategies. The irregularity in these segmentation outcomes raises concerns about the reproducibility of the results, a challenge that we aim to address in this work.

In our study, we employ the mass balance of a metamorphic reaction as an internal standard to verify segmentation accuracy and shed light on the advantages of deep learning approaches, particularly their capacity to efficiently process expansive datasets. Our methodology utilises deep learning to achieve accurate segmentation of time-resolved volumetric images of the gypsum dehydration reaction, a process that traditional segmentation techniques have struggled with due to poor contrast between reactants and products. We utilise a 2D U-net architecture for segmentation and introduce machine learning-obtained labelled data (specifically, from random forest classification) as an innovative solution to the limitations of training data obtained from imaging. The deep learning algorithm we developed has demonstrated remarkable resilience, consistently segmenting volume phases across all experiments. Furthermore, our trained neural network exhibits impressively short run times on a standard workstation equipped with a Graphic Processing Unit (GPU). To evaluate the precision of our workflow, we compared the theoretical and measured molar evolution of gypsum to bassanite during dehydration. The errors between the predicted and segmented volumes in all time-series experiments fell within the 2% confidence intervals of the theoretical curves, affirming the accuracy of our methodology. We also compared the results obtained by the proposed method with standard segmentation methods and found a significant improvement in precision and accuracy of segmented volumes. This makes the segmented CT images suited for extracting quantitative data, such as variations in mineral growth rate and pore size during the reaction.

https://doi.org/10.5194/egusphere-2023-1819
Preprint. Discussion started: 14 August 2023

In this work, we introduce a distinctive approach by using an internal standard to validate the accuracy of a segmentation model, demonstrating its potential as a robust and reliable method for image segmentation in this field. This ability to measure the volumetric evolution during a reaction with precision paves the way for advanced modelling and verification of the physical properties of rock materials, particularly those involved in tectono-metamorphic processes. Our work underscores the promise of deep learning approaches in elevating the quality and reproducibility of research in the geosciences.

1 Introduction

Time-resolved (4D) operando experiments in μCT scanners have emerged as a promising way of studying solid-state reactions offering unprecedented insight into mineral phases and volume changes and the method is becoming a technique of choice for many geoscience problems because it provides information about both the spatial and temporal evolution of the microstructure of a sample. This technique can achieve a range of resolutions, with voxel sizes from millimetres to hundreds of nanometres. Underpinning any usefulness of these new insights is the accurate segmentation of individual phases into three-dimensional (3D) representations across often large datasets; once different phases are segmented and labelled, they directly aid in a quantitative understanding of all types of solid-state mineral reactions (metasomatic, diagenetic, metamorphic, and physico-chemical alteration) (Fusseis et al. , 2014).

For the accurate quantification of the various phase components and evolution of minerals from 4D μCT data, semantic segmentation needs to be accomplished. Semantic segmentation refers to labelling individual pixels of an image to a corresponding classification. Image segmentation has long played a pivotal role in the quantitative analysis of digital representations of geological materials and there is now a wealth of methods available (Reinhardt et al., 2022). However, not all segmentation workflows can effectively track a process in space and time across different samples and acquisition conditions, as is needed in the case of *in-situ*, or *operando* time-resolved X-ray microtomography studies. For instance, while standard histogram segmentation can be consistently applied to a single time-step, it may not be easily transferable between different samples undergoing solid state transformation (Andrew, 2018). More advanced machine learning techniques have been used successfully on a range of geoscience problems and offer better portability and applicability compared to histogram segmentation (e.g. for solid state reactions (Marti et al., 2021); crack detection (Cartwright-Taylor et al., 2022; Lee et al., 2022; Reinhardt et al., 2022); and one/two-phase flow experiments (Phillips et al., 2021), but they also still fall short in achieving complete portability between various time-steps. While deep learning methods show promising potential for tackling the challenges in image segmentation of high-resolution, time-series datasets, they still need refinement for optimal performance.

Deep learning algorithms are gaining popularity for analysing microstructures in biological and medical sciences (Renard et al., 2020), in engineering materials (e.g., Müller et al., 2021; Allen et al., 2022) and for the segmentation of deforming and reacting porous rock materials (Da Wang et al., 2021). However, regardless of their scientific domains, most studies

https://doi.org/10.5194/egusphere-2023-1819
Preprint. Discussion started: 14 August 2023

focus on two-component systems, void and solid material classifications. In addition, some algorithms still rely heavily on adaptive filtering and simple thresholding operations (Phan, 2021). This is a limitation because greyscale images contain limited information and this restricts how effectively a deep learning algorithm can perform, regardless of its complexity. This is most clearly seen in data that contain low contrast phases. Moreover, as grayscale image inputs evolve in time-series datasets the usefulness of any thresholds chosen is undermined. Greater insight into microstructural changes can only be gained through the full segmentation of all mineral components based on grayscale and other considerations, like for example component morphology. This outlines a clear need for deep learning workflows to be further explored and optimised so that they can be better exploited in geosciences.

In this paper, we explore the use of supervised deep learning to segment 4D synchrotron-based μCT datasets of dehydrating Volterra Alabaster (Fig. 1). We employ a 2D U-Net architecture (Ronneberger et al., 2015) and demonstrate its capability to accurately segment the data into four phases: gypsum, bassanite, celestite, and pores. This model dehydration reaction has been monitored during experiments under different stress and pore fluid pressure conditions (Gilgannon et al.,, 2023). The data used encompass numerous challenges encountered in volumetric image segmentation of complex materials, including multiple heterogeneous material phases with feature sizes ranging from hundreds of nanometres to micrometres, low contrast between phases, and a relatively rapid evolution. We demonstrate that these factors make segmentation using standard approaches difficult. We quantitatively compare outputs of the deep learning architecture to optimise its use and for the first time show how the accuracy of segmentations can be checked with an internal standard given by the chemistry of the system. Ultimately, we find that the use of a random forest classifier to produce the 'ground truth' to the training of the deep learning architecture improves the predictive abilities of the algorithm. While the random forest algorithm initially can effectively segment features of interest in our dataset, its capability for generalisation to new, unseen data is limited (Rezaee et al., 2018). The inclusion of the deep learning step enhances the generalisation capability of our workflow. This results significantly improved accuracy and validity of the segmentation and labelling of the μCT data during the solid state reaction of gypsum to bassanite and pore space. We believe that this work demonstrates the potential of deep learning for volumetric image segmentation of complex materials. The method is generic and can be applied to other geoscience problems.

2 Gypsum Dehydration as an Example of a Complex Segmentation Problem

2.1 The Gypsum Dehydration System and Experimental Set-up

Gypsum dehydration is used as a model dehydration for many prograde metamorphic reactions in collisional tectonic settings. The physical boundary conditions make it amenable for laboratory studies and thus a system of choice to investigate complex geological problems in time-resolved μCT in-situ experiments.

Volterra Alabaster is a rock that is mainly (>90 %) composed of gypsum ($CaSO_4 \cdot 2H_2O$) and celestite ($SrSO_4$) and when temperature is increased the gypsum dehydrates to produce bassanite ($CaSO_4 \cdot \frac{1}{2}H_2O$), porosity and water. At the same time, celestite remains stable during gypsum breakdown and is unaffected by the dehydration. The dehydration of gypsum results in a 29% reduction in solid molar volume, and an 8% excess volume of water. Experiments were performed with the x-ray

https://doi.org/10.5194/egusphere-2023-1819
Preprint. Discussion started: 14 August 2023

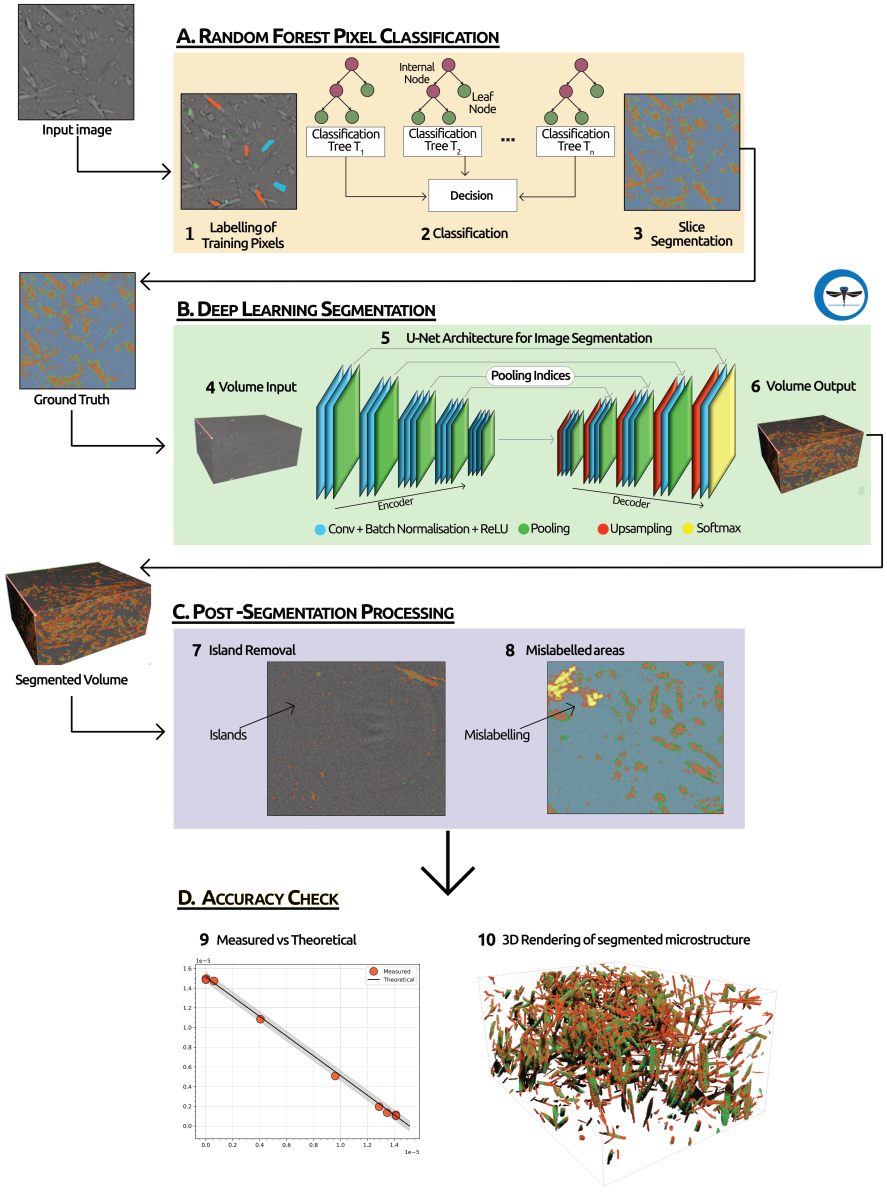

Figure 1. Workflow used for image segmentation. a) The first step involves labelling the different phases (i.e., Gypsum, Bassanite, Pores, Celestite) over a few (13) slices in the volume and then applying a Random Forest (RF) pixel classification. b) The second step of our segmentation process involves using the output of the RF as ground truth and then running a 2D U-net deep learning algorithm over the whole selected volume. c) In the third step we apply a series of post-segmentation routines to clean the data set from possible segmentation errors. d) In the final step we quantitatively evaluate the overall performances of the trained deep learning network by comparing the theoretical and measured molar evolution of gypsum to bassanite during the dehydration.

https://doi.org/10.5194/egusphere-2023-1819
Preprint. Discussion started: 14 August 2023

transparent triaxial rig Mjölnir (Butler et al., 2020) at the TOMCAT beamline of the Swiss Light Source synchrotron. All of our experiments were performed at the same confining pressure (P_c) of 20 MPa and a pore fluid pressure (P_f) varying between 1 and 5 MPa. The experiments followed the same temperature path, with a maximum temperature 124.5 $<$ T $<$ 126.9 °C. We systematically varied the differential stress in each experiment to capture its effect (σ_{diff} = 0; 16.1; 27.9 MPa, see Gilgannon et al., (2023)). For the work presented in this paper, however, we focus on data from two specific experiments: (a) a sample, VA17, where the principal stress is radial (with σ_{diff} = 11.3 MPa), and (b) another sample, VA19, where the principal stress is vertical (with σ_{diff} = 16.1 MPa). Time-resolved (4D) synchrotron microtomography (μCT) datasets were acquired during the experiments at SLS TOMCAT beamline using a pink beam with an energy peak at 27 KeV. For each μCT dataset, 1500 radiographs were collected over 180° rotation in 2-4 s. The resulting radiographs had a voxel size of 2.753 μm, and the resulting 3D μCT datasets had a size of 2016 $\times$ 2016 $\times$ 2016 voxels. The frequency rate of the tomoscopy was set to 60 seconds, and the experiments ran over 150-314 minutes, resulting in 2.5 TB of data to be analysed. More details on the experiments can be found in Gilgannon et al., (2023).

2.2 Challenges of Segmenting Dehydrating Gypsum during Operando X-ray Microtomographies

It is clear from Figure 2 that microstructural changes during the experiment can be readily identified by the human eye. However, it is also apparent from the evolving histograms in Figure 2a that accurate segmentation of the four phases of interest (i.e., gypsum, bassanite, pores, and celestite) cannot rely on simple histogram thresholding segmentation. Each 4D μCT dataset is extensive, containing more than 100 scans, each ranging from ∼5 to ∼15 GB of reconstructed data, depending on the scanning parameters. As the histograms of individual scans are clearly different for different time steps (Fig. 2a), an automated segmentation of the evolving volumes based on a single histogram would yield inaccurate results, necessitating manual segmentation and the explicit selection of thresholds for each tomogram and for each experiment. This laborious process inhibits the efficient analysis of large 4D datasets but also misses basic standards for reproducibility. This is further complicated by the fact that all synchrotron μCT images have a symmetrical vertical gradient in noise through the sample, first decaying and then increasing, which renders the application of a single set of thresholds even to a single μCT dataset problematic. Additionally, the homogeneity of the unreacted starting material intensifies artefacts such as rings which are problematic to handle for segmentation algorithms that are based solely on grayscale thresholds. As noted above, the human eye can distinguish different phases in the data and this suggests that a learning-based approach to semantic segmentation would be applicable to the dataset. It is becoming evident that we may also require information beyond grayscale values, such as the geometry of the feature of interest, for successful segmentation.

2.3 A Segmentation Workflow with Internal Standards

To accurately segment large datasets of dehydrating gypsum samples, we used deep learning algorithms, which have entered the field of volumetric image segmentation through the implementation of convolutional neural networks (CNN). For this work, we used a specific implementation of 2D U-Net available in the Dragonfly™ software. This implementation has performed well on μCT images from fibre-reinforced ceramic composites Badran et al. (2020). Our dehydrating gypsum datasets are

https://doi.org/10.5194/egusphere-2023-1819
Preprint. Discussion started: 14 August 2023

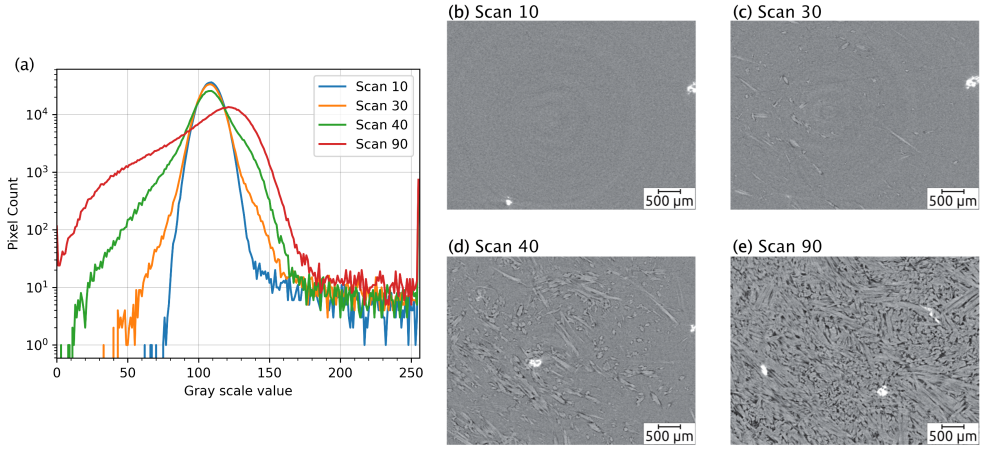

Figure 2. Variation in grayscale intensity values with reaction progress. (a) Grayscale histograms of a series of tomographic slices captured at different stages during the reaction. Each line represents the histogram of an individual image, illustrating the changes in grayscale intensity value distribution across the time steps. This variation complicates the use of grayscale thresholding as input for the deep learning model. (b)-(e) Display the images corresponding to the histograms in (a), depicting the reaction progression from an early stage (b) to the final product (e).

comparable in terms of grayscale value contrast and the number of distinguishable material phases. Dragonfly runs locally on a workstation and allows for the creation of training data and training of a CNN segmentation model. Once the CNN is trained, the model can be generalised and offers the advantage of being flexible and being straightforward to apply on similar datasets.

To train the network, we selected 13 of the 2016 horizontal (XY) virtual slices from the synchrotron CT scan of sample VA19 time step 40 as "input" images. This specific time-step was chosen because it has sufficient volume of each phase we aim to segment; in images derived from either early or late steps of the experiment, the volume of at least one of the phases would be insufficient to achieve automatic segmentation. We tested the role of ground truth data (i.e., the correct segmentation of an image) in achieving the best results by comparing a histogram thresholding segmentation with a random forest classifier.

Choosing the best training neural network architecture and tuning the network (hyper-)parameters requires time and some knowledge of neural network architecture. However once the best configuration is set up the application of the model is nearly effortless. Network (hyper-)parameters that need to be chosen include: (i) a "patch size" – in the training stage the images are split into a set of smaller 2D square patches that capture the features of interest in the image; (ii) a "stride ratio" – which defines the position of the neighbouring patches (at a value of '1.0', there will be no overlap between patches and they will be extracted sequentially one after another; at a value of '0.5', there will be a 50% overlap); (iii) a "batch size" – which defines the number of patches evaluated in each batch prior to updating the network model; (iv) the number of epochs – an epoch indicates a training iteration, involving a pass over all batches of the training set; (v) selection of a loss function to evaluate how far the output of the CNN model deviates from the ground truth and an optimisation algorithm to find optimal weights for the coefficients of the CNN.

https://doi.org/10.5194/egusphere-2023-1819
Preprint. Discussion started: 14 August 2023

We trained the different networks by varying the (hyper-)parameter settings to see which setting results in a measurable improvement to model performance. For all the tested strategies, twenty percent segmented data serves as a "validation set" and is otherwise not used during training. A loss function was used to evaluate the training progress. The U-Net deep learning architecture was trained for a maximum of 100 epochs, stopping when no further improvement of loss was observed.

To demonstrate the accuracy of the segmentations, we devised an additional quality check consisting of comparing the output volumes of phases to their predicted values given by the mass balance of the reaction (see section 3.7). This internal standard allows us to objectively assess the effectiveness of the application of deep learning to time series data sets that contain low contrast phases.

3 Influence of Training Data

3.1 Input Data for the Deep Learning Convolutional Neural Network

Convolutional neural networks (CNN) are a special class of deep learning algorithms where one or more layers of the network perform convolution operations (Fig. 1). The specific convolution kernels are not programmed but are learned from the input data by the deep-learning engine to extract relevant features of an image that become useful discriminators in segmenting complex pixel classes – in our case mineral phases – in the images. The CNN architecture (Fig. 1b) can be thought of as a formula of linear weights applied to the image pixel intensities, often combined through multiple network layers in a nonlinear fashion. The coefficients encoded in the neural network itself are learned from training data that couples example "input" images (i.e., the raw un-segmented image) with example "output" images (i.e., ground truth). The iterative process of learning the weights that can reliably transform input into output images is termed training and is the most computationally demanding phase of the deep learning cycle. In order for a deep learning model to perform the segmentation of the different classes contained within the image, it requires a set of data which are typically created by manually annotating each pixel in the image with its corresponding semantic label (Fig. 1a). This set of labelled pixels forms the so-called "ground truth". Ground truth is an essential part of training deep learning models as it represents the target to learn towards and should ensure that the model is learning to segment images in a meaningful way. In our case, the ground truth images are a selection of image slices that have been previously segmented to assign each pixel to a specific mineral or pore phase. The trained model can then, not only automatically segment the remaining unsegmented image slices within a single μCT volume, but also unseen data - i.e., other volumes within the same time series of the training set as well as image volumes obtained during other experimental time series.

Initial image classification to create the ground truth set can be made with various levels of information, like for example (i) histogram thresholding – low information levels – or (ii) Greyscale information alone, as provided by histogram thresholding, would express gradients in a field of values for which boundaries between phases are diffuse. In contrast, ground truth classified using a machine learning algorithm provides discrete transitions between phases and better information about features like phase morphology. Before finding the best workflow to segment our image volumes, we tested several combinations of ground truth input and CNN parameters until the quality of output images on unseen data was adequate, with reasonable training time.

https://doi.org/10.5194/egusphere-2023-1819
Preprint. Discussion started: 14 August 2023

3.2 Histogram Thresholding as Training Data

Figure 3a shows an example of data containing the four phases as segmented by different networks trained with different ground truth data. The corresponding bracketed threshold values for the four phases are shown in Figure 3b. Figure 3c shows the output produced by a neural network trained using a ground truth set of images labelled solely by manual histogram thresholding. This input failed to reliably classify the bassanite needles, it often failed to segment pores, and it failed to accurately segment celestite, often confusing it for bassanite. This can be improved upon by using manual histogram thresholding with the application of data augmentations within the neural network model (Fig. 3d). The training data were subjected to data augmentation based on the basic image manipulations (i.e., flip horizontally, flip vertically, rotate, shear, and scale). Specifically, we octupled the input data in order to render the neural network more robust, while at the same time compensating for deliberately using a small input dataset. This strategy allowed us to increase the network's ability to generalise while decreasing the potential danger of overfitting (Shorten & Khoshgoftaar, 2019). By tuning the different CNN (hyper-)parameters and including augmented data, we improved the overall performance of the network (Fig. 3d). However, the final segmentation still lacked accuracy: it can be seen that errors remained, for example the celestite was still identified as bassanite (Fig. 3d). More importantly, this deep neural network model struggled when it was applied to new and more complex datasets: such as in the early and final stages of the reaction (where one of the two main phases was scarce, or absent).

3.3 Random Forest Classifier as Training Data

In contrast, the use of ground truth data classified with a random forest classifier plus data augmentation performed exceptionally well and visually captured more of the features of the microstructure correctly (Fig. 3e).

A random forest classifier (Fig. 1) comprises numerous decision trees, each contributing a vote toward the class prediction for every voxel (more details on the random forest classifier are available in the Additional Information). The class receiving the most votes is allocated to the respective voxel. Filters are applied to the input images, generating filtered images that serve as features (Reinhardt et al., 2022). These features enable the classifier to differentiate between phases in the dataset. In this work, the random forest classifier was pre-set with morphological filters, 3x3 neighbour filter and Gaussian filter to perform identification of the phases in the training set. Using a random forest classifier for setting up the ground truth dataset also enabled training the deep learning model based on the shape of the objects. This offered a significant progress from manual thresholding segmentation.

3.4 Optimising the Deep Learning Models

Model parameters can be optimised to improve the deep learning network. We systematically monitored the performance of each tested deep learning model both during training and testing. The results of this systematic testing are visualised in Figure 4 and synthesised in Table 1. The quantitative comparison of types of ground truth data and the variations of model (hyper-)parameters provides a solid base for discussing the advantages of the workflow that is presented here and how it is transferable to other geoscience data.

https://doi.org/10.5194/egusphere-2023-1819
Preprint. Discussion started: 14 August 2023

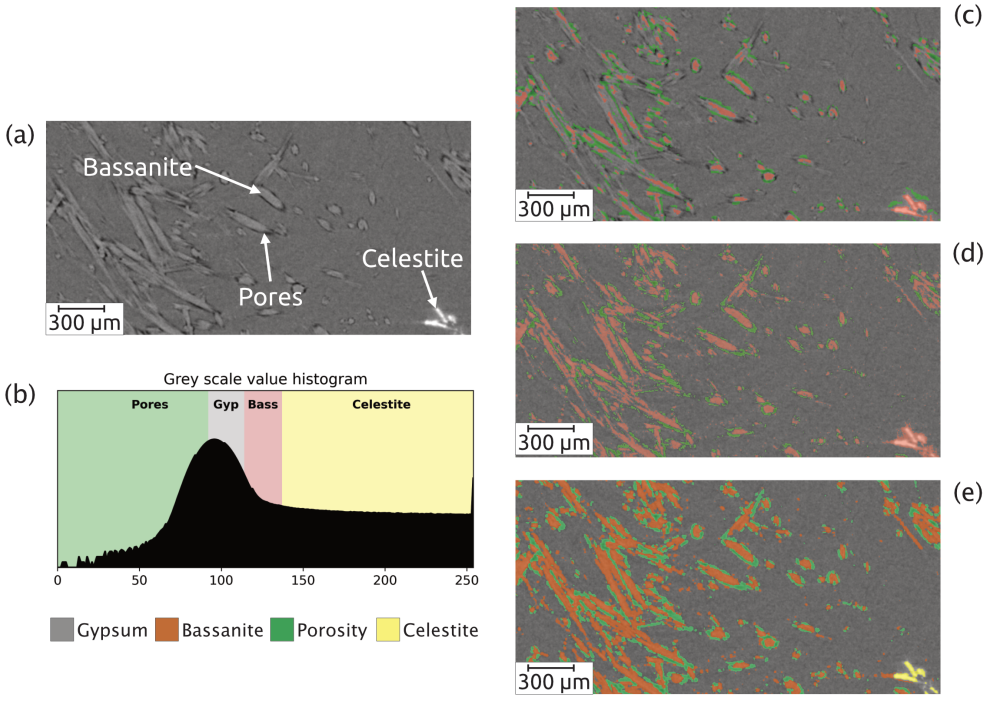

Figure 3. Challenges of the segmentation. (a) Portion of a horizontal slice of the raw μCT image showing the relative low contrast between gypsum and bassanite. (b) threshold values for the four phases present in the image. Two main phases – Gypsum and Bassanite – are difficult to split up accurately into two classes by the deep learning algorithm with (c) no data augmentation but are better segmented when using data augmentation (d) but still showing evident artefacts in the segmentation. Both cases (in c and d) struggle in separating the celestite phase which is wrongly classified as bassanite. (e) the segmentation results using Random Forest classifier as input into the deep learning algorithm.

For an objective quantitative comparison of different deep learning network models, we tracked the performance of each model during training using a loss function to measure the error between the neural network's prediction and the corresponding ground truth; the error was then used to update the model parameters. Figure 4a shows that with ground truth data derived from random forest classification we obtained the lowest validation errors for all tested networks. Compared to other models, the random forest model reached low values of loss already after 5 epochs. After this minimum, the error kept oscillating within the neighbourhood of its lowest value until the maximum 100 epochs were reached. This indicates that the overfitting risk is minimal.

We evaluate segmentation quality according to eight standard evaluation metrics based on overlap and similarity criteria (Taha & Hanbury, 2015; Müller et al., 2021), where the deep learning-based segmentations are compared to the corresponding ground truths. Here we focus on the Dice coefficient, however, a full picture of all calculated metrics can be found in the Supplementary Information. The Dice coefficient scores were used to evaluate and compare the segmentations resulting from

https://doi.org/10.5194/egusphere-2023-1819
Preprint. Discussion started: 14 August 2023

Table 1. Evaluation of the segmentation models based on Dice Coefficient scores

Name Model	Model A	Model B	Model C	Model D	Model E	Model RF
Ground Truth*	HT	HT	HT	HT	HT	RFC
Data Augmentation	NO	NO	NO	YES	YES	YES
Training Parameters*	P=16 S=1 B=64	P=16 S=0.5 B=128	P=16 S=1 B=128	P=16 S=1 B=64	P=16 S=1 B=128	P=16 S=1 B=512
Average DICE	0.013	0.021	0.095	0.088	0.964	0.961
Gypsum DICE	0.035	0.041	0.651	0.546	0.983	0.979
Bassanite DICE	0.001	0	0.077	0	0.845	0.923
Pore space DICE	0.019	0.024	0.122	0.160	0.957	0.924
Celestite DICE	0	0	0	0	0	0.827

*Abbreviations: HT = Histogram Thresholding; RFC = Random Forest Classifier; P = Patch Size; S = Stride Ratio; B = Batch Size

neural network models trained on (i) histogram thresholding ground truth data (models A, B and C), (ii) histogram thresholding with augmented ground truth data (models D and E), and (iii) random forest classified ground truth data with augmentation (model RF). The Dice coefficient (DICE) is the normalised overlap of pixels in the segmentation and the corresponding ground truth of a given phase. A DICE score of 0 means that there is no overlap between segmentation and ground truth, while a DICE score of 1 indicates perfect overlap. In addition, to the direct comparison between automatic and ground truth segmentations, it is common to use the DICE to measure reproducibility (repeatability) of a trained neural network segmentation algorithm (Taha & Hanbury, 2015).

For the networks trained using histogram thresholding the average DICE varies between 0.01 and 0.98, it increases when data augmentation is used during training; For the network trained using a Random Forest classifier, the DICE score is also 0.98 (Table 1). From the error curves and the DICE plots, it is clear that the inclusion of augmented data into the histogram threshold ground truth (as seen in model D and E in Fig. 4 and Table 1) improved the overall performance of the neural network model compared to the models which did not (model A, B and C in Fig. 4). The DICE scores for each segmented phase show similar trends, on average improving for data augmented models. However, it was only the model trained using a ground truth from a random forest classifier that produced scores for all four phases. This includes the celestite phase, which was entirely absent in the results from the other models (Fig. 4).

All these results show that using a random forest classifier pre-classified ground truth data clearly outperforms a ground truth obtained via simple grayscale histogram thresholding regardless of the optimisation of parameters.

3.5 Applying the deep learning segmentation

After the training stage, the model was applied to a larger sub-volume (400 consecutive slices, ~250 MB) from the same scan used during training of the deep learning algorithm (i.e., VA19 time step 40). The model succeeded in correctly segmenting all four phases in about 7 minutes (each volume composed of 400 slices, with 250 MB each volume) using a computer with 256GB of RAM, an Intel Xeon Eighteen-Core Processor, and an NVIDIA Quadro RTX 5000 16GB GPU. Typical results are

https://doi.org/10.5194/egusphere-2023-1819
Preprint. Discussion started: 14 August 2023

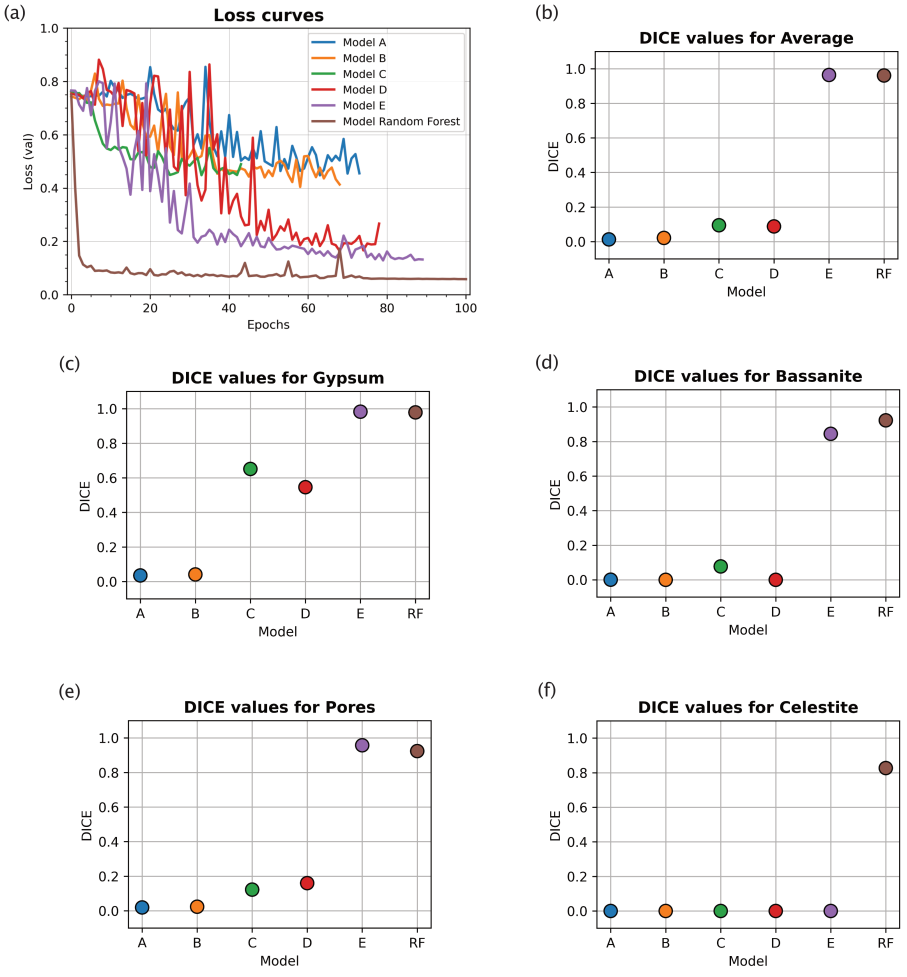

Figure 4. Error curves and evaluation metrics for the image segmentation models (see Table 1 for details on the models). (a) Error curves comparison for different loss values for the tested models. The neural network model trained with a random forest Classifier input outperformed the other models which used manually thresholded inputs. Plots for Dice coefficient (DICE) metric (colour coded as in (a)), show the predictive performances of each trained model for both the average volume (in b) and each separate phase: gypsum (c), bassanite (d), pores (e), and celestite (f). Clearly, The model trained using a random forest classifier input demonstrates superior performance to other models. See Table 1 for details on the training parameters for the different models.

shown in Figure 5, which compares equivalent horizontal and vertical slices from the unprocessed and segmented CT images. This comparison indicates that gypsum, bassanite, porosity and celestite are clearly labelled, even in the portions showing ring artefacts that can mask the true grayscale values (Figure 5c, d). Importantly, the combination of random forest-based ground truth and deep learning segmentation ensures that the ring artefacts are not mislabelled as actual phases, a problem that

https://doi.org/10.5194/egusphere-2023-1819
Preprint. Discussion started: 14 August 2023

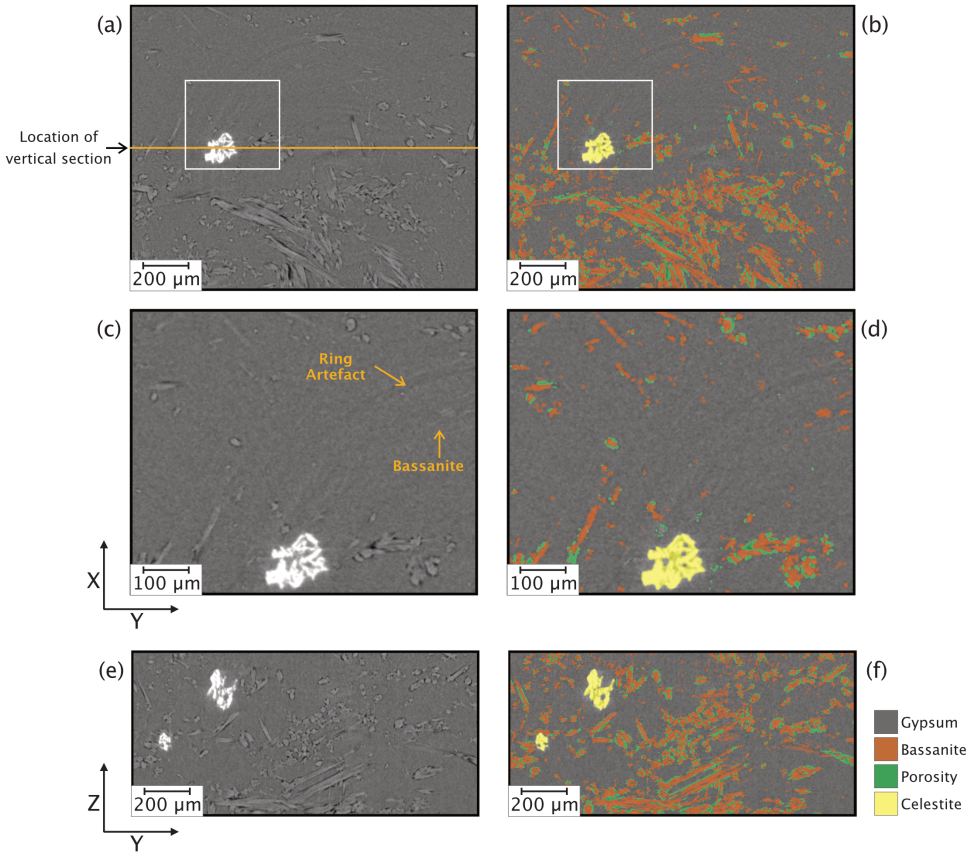

Figure 5. Horizontal (XY) and vertical (YZ) slices of the μCT images before (a, c, e) and after (b, d, f) deep learning segmentation performed using a model trained with Random Forest classified images. Images in (c) and (d) are closer captions of areas indicated with white boxes in (a) and (b).

frequently arises with manual histogram thresholding. This consequently prevents the creation of fictitious phases when there is none. Comparison of the vertical (YZ) sections (Figure 5e, f), where the unprocessed slice clearly shows all four phases, qualitatively indicates that the accuracy of the segmentation is high.

3.6 Post-segmentation processing

Once the data volume is segmented by the trained deep learning model, we apply a series of post-segmentation routines to clean the data set from segmentation errors, which is necessary primarily on data acquired early in the experiment when the contrast in the sample was low. These routines involve removal of isolated clusters of erroneously labelled pixels and deletion of areas labelled as bassanite around celestite aggregates. The first routine is implemented using the "remove island" tool in Dragonfly™, targeting pixels misinterpreted as bassanite and porosity. The size of clusters to be removed is fixed in all

https://doi.org/10.5194/egusphere-2023-1819
Preprint. Discussion started: 14 August 2023

volumes of the time series and also through the different scanned samples: 100 pixels and 8 pixels for bassanite and porosity, respectively. The application of this routine is much more frequent at early stages of the dehydration process when the majority of the volume is still represented by the gypsum phase: the lack of contrast between phases leads to very noisy slices (i.e., speckled in appearance). The second routine involves the deletion of mislabelled areas around celestite. This procedure is most accurate and fast if conducted through visual inspection and manual corrections.

3.7 Understanding the accuracy of the segmentation

Time-resolved μCT data offer the opportunity to quantify evolving volumes in a sample and thereby the rates of a process, whereby the accuracy of the quantification hinges on the accuracy of the volumetric segmentation. The accuracy of our deep-learning segmentation method itself is contingent upon three potential sources of error. The first pertains to the quality of the original CT image data, influenced by factors such as image resolution, noise, and potential artefacts. In our work, this source of error had minor impacts on the segmentation of bassanite and pores and is primarily restricted to the early stages of the dehydration process when the dominant presence of a single mineral phase, gypsum, led to noisy slices and enhanced ring artefacts. A second potential error source lies in the initial segmentation used to establish the target image. The initial segmentation is arguably the most laborious and time-consuming step, with some level of error inevitable during the labelling of slices and assignment of pixels to specific phases, particularly at phase boundaries. These issues, however, have minimal impact on the final trained model as they generally occur at isolated pixels (Badran et al., 2020). The third potential source of error is mislabelling of pixels during the deep-learning segmentation stage, attributable to limitations in the accuracy of the trained model. While the error rate typically decreases with an increase in the number of training images and iterations, overfeeding the training network can lead to overfitting, which can in turn degrade performance when segmenting unseen images. However, we showed (Figs. 3 and 4; Table 1) that augmenting data significantly reduced both mislabelling and overfitting during the training step of the neural network.

The quantification of a metamorphic reaction rate from 4D μCT data hinges on the accurate tracking of the evolution of reacting and emerging phases. To independently ascertain the accuracy of the chosen deep learning model, we compared the theoretical and measured (i.e. segmented) molar volumetric evolution of gypsum to bassanite during the dehydration reaction. To our best knowledge, this represents the first application of an internal standard to unambiguously measure the accuracy of a segmentation model.

For the case studied here, where no irreversible compaction occurred in the samples during the experiments (Gilgannon et al.,, 2023), we can use the theoretical molar evolution during the dehydration of gypsum to bassanite to calculate the amounts of gypsum, bassanite, and water produced during the dehydration reaction and the stoichiometric ratios between them. Gypsum has two water molecules per formula unit, while bassanite has only half of a water molecule per formula unit. Hence, during the dehydration process, the molar ratio of water molecules to calcium sulphate molecules decreases. The chemical equation for the dehydration of gypsum to bassanite is:

https://doi.org/10.5194/egusphere-2023-1819
Preprint. Discussion started: 14 August 2023

$$CaSO_4 \cdot 2H_2O \rightarrow CaSO_4 \cdot 0.5H_2O + 1.5H_2O \quad \text{(R1)}$$

where one mole of gypsum ($CaSO_4 \cdot 2H_2O$) gives one mole of bassanite ($CaSO_4 \cdot 0.5H_2O$) and 1.5 water molecules.

Knowing the initial volume of gypsum in the sample and its density (2310 kg/m^3), we can simply calculate its mass and the corresponding molar quantity. From this, we can compute the theoretical amount of bassanite produced from gypsum at every reaction step. Given that the density of bassanite is 2731 kg/m^3, we can use the molar mass to convert the produced moles of bassanite into volume. A plot of the moles of reactants versus products is a y=-x graph consistent with the 1:1 stoichiometric ratio of reaction (Eq. 1). For the 1:1 gypsum to bassanite reaction, the slope is -1 (solid black lines in Figs. 6a and b with grey-shaded 2% confidence intervals). The segmentation, which provides a volume of bassanite and gypsum at each step, can be represented and compared to the theoretical case. This graphical method forms the basis of the theoretical dehydration curve against which we compare the segmented volumes. Examples of this comparison are shown in Figure 6, where we present dehydration evolution paths for a sample under radial stress (VA17, Fig. 6a) and an axial deviatoric stress sample (VA19, Fig. 6c). In both examples, the curves for theoretical and measured molar volumes follow the same trend, with fitting parameters showing close equivalence.

This comparison shows that our segmentation workflow produces highly accurate volume fractions for each phase. All fractions fall within the <5% confidence intervals of the theoretical curve (Fig. 6). For a more comprehensive evaluation of the method, a comparison was made between the novel integrated workflow (Fig. 7a), and traditional manual histogram thresholding. This comparison was applied to a selection of volumes in the time series (Fig. 7b). The manual thresholding method, which incorporates basic pre-processing steps (including “despeckle” and “non-local means” with sigma = 5, smoothing = 1) displayed significant shortcomings. It resulted in a severe underestimation of the reaction extent and the inadvertent ’creation’ of celestite. Contrarily, the proposed workflow (Fig. 7a) significantly outperforms the traditional approach (Fig. 7b).

Due to its demonstrable accuracy, the segmentation output is well-suited for extracting quantitative information, such as mineral growth rates and variations in pore size during the dehydration reactions. Our segmentation method enables the quantification of relative accuracy, allowing for the propagation of errors in any derived and quantified parameters. This advance represents a significant step towards interpreting results and establishing their significance, as confidence intervals are often absent in studies using manual thresholding.

4 Discussion and implications

The application of deep learning to time-resolved micro-CT imaging offers a new tool for geoscientists studying rock deformation, metamorphic processes, and fluid-rock interactions. We successfully leveraged optimised deep learning methods to perform reliable and efficient segmentation of time-resolved volumetric images during the gypsum dehydration reaction. The approach outlined here not only streamlines data analysis by swiftly processing large datasets, but also enhances confidence in the robustness of results by ensuring high segmentation accuracy. Importantly, this accuracy is established in a robust way

https://doi.org/10.5194/egusphere-2023-1819
Preprint. Discussion started: 14 August 2023

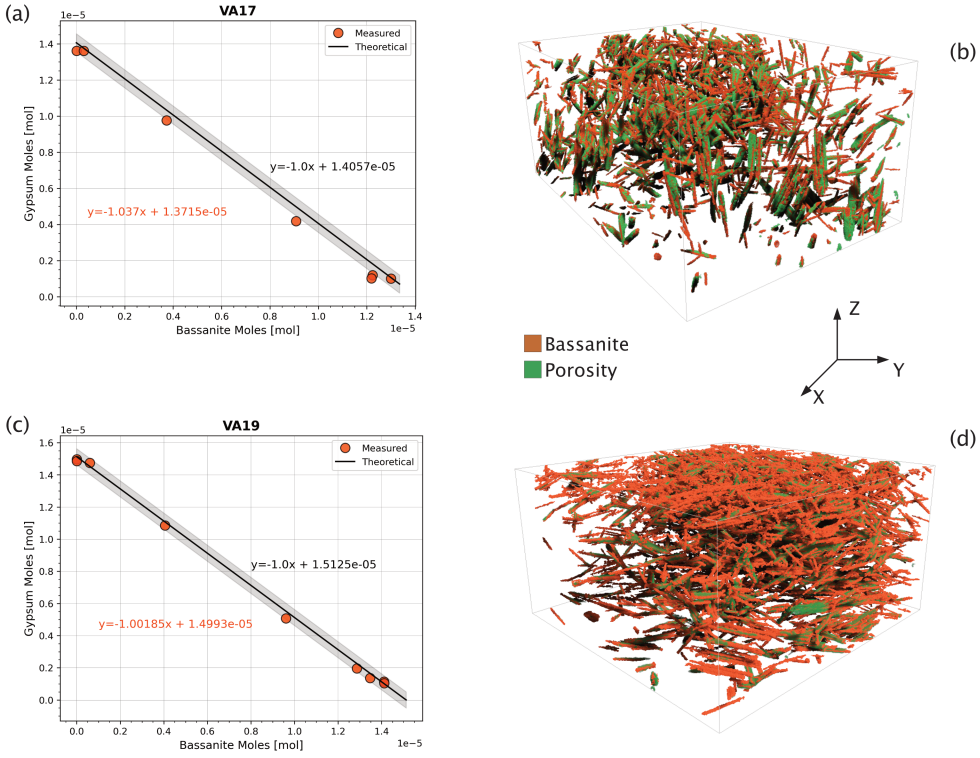

Figure 6. Theoretical versus measured. Comparison between the theoretical and the measured molar ratio of gypsum and bassanite during dehydration. Here, we plot the evolving molar ratio of gypsum to bassanite during dehydration for (a) sample VA17 which experienced radial stress and (c) sample VA19 experiencing axial stress (see Figure 5 for reference). Shaded grey areas are 2% confidence intervals of the theoretical curve. On the right-hand side, 3D renderings of bassanite crystals and pores in two samples' sub-volumes reacting at two different stress conditions. In (b) VA17 reaction principal stress (σ_{max}) is axial (i.e. parallel to Z), while in (d) VA19 the principal applied stress is radial (in the X-Y plane). Heights of both boxes are 1.5 mm.

due to the three-component system under study: gypsum, bassanite, and water. We ascertained the accuracy of the chosen deep learning model we compared the theoretical and measured molar evolution of gypsum to bassanite during dehydration. This approach defines an internal standard, verifying that the segmentation method accurately captures the mineralogical changes occurring within the rock samples. Importantly, the robustness of this validation is based on the three-component nature of the system—gypsum, bassanite, and water (imaged as porosity the μCT data)—allowing for a non-circular and independent verification of our method's effectiveness. By harnessing the power of deep learning for image segmentation, we can extract more nuanced and precise information from μCT imaging datasets. This will enhance our understanding of geological processes and contribute to more accurate models of rock behaviour under different physical conditions.

https://doi.org/10.5194/egusphere-2023-1819
Preprint. Discussion started: 14 August 2023

4.1 Comparison with other segmentation approaches

Accurate segmentation has long been a challenge across various scientific domains, from medical CT imaging to material sciences and engineering (Withers et al., 2021). Global segmentation methods, such as manual histogram thresholding, have long been go-to solutions for the segmentation of X-ray μCT tomographic images. However, three considerable drawbacks persist: (i) significant time commitment required, (ii) global techniques ignore local context and thus have an intrinsic potential for misclassification, and, therefore, (iii) the potential for compromised reproducibility (Andrew, 2018) As μCT imaging technologies evolve, resulting in larger datasets, the scalability and efficiency of manual segmentation methods become increasingly challenging (Da Wang et al., 2021). Herein lies the risk of jeopardising reproducibility, defined as the ability to consistently obtain similar results across multiple measurements using the same methodology (Renard et al., 2020).

Machine and deep learning segmentation strategies form promising alternatives for automatic segmentation, optimising parameters for high accuracy performance on the training dataset and ensuring effective generalisation to other datasets within the same problem class. However, transitioning towards automatic segmentation, while promising, is not trivial. Successful automation of segmentation methods still requires an initial investment of time and resources for skill acquisition and understanding needed to fine-tune models and adapt the workflow to the specific dataset at hand.

A good example of this is the intrinsic dependency of deep learning segmentation on the ground truth data input and the selection of hyperparameters during the training process. If the initial segmentation—which forms the ground truth—is not meticulously executed, this could lead to subpar results. These could manifest as minor differences when compared with the ground truth, creating a misleading perception of accurate segmentation. Given these potential pitfalls, independent verification of segmentation results appears to be a preferable approach.

In fields such as medical imaging and material science a common strategy to ensure reliability and accuracy of segmentations is the use of external calibration techniques, which involve the use of phantoms with known dimensions and/or compositions as benchmarks (Adams , 2009; Kruth et al., 2011). These external standards aid in the assurance of measurement accuracy (Withers et al., 2021). These external standards aid in the assurance of measurement accuracy [Writers et al., 2021]. However, these calibration techniques are not without limitations. One major challenge lies in partial volume effects, which occur when the volume of interest encompasses more than one type of material. The CT values measured in these regions do not correspond to a single material type, but rather are a weighted average of the different types present (Kruth et al., 2011; Sokac et al., 2020). Solutions have been proposed and often require complementary techniques (such as using tactile, optical sensors) to calibrate measurements derived from CT data (Torralba et al., 2018). Furthermore, the use of phantoms can result in difficulties during sample preparation (such as staining, sample chemistry/structure modification to include the standard) which, in turn, can alter the general output of the segmentation.

In line with the efforts to enhance the accuracy and reproducibility of CT image-based measurements, our approach leverages on an *a-priori* knowledge of the chemical reactions involved in the dehydration process, therefore establishing a framework for assessing the accuracy of the data extracted from the μCT images. This internal validation approach offers a robust and consistent means of assessing the reliability of our segmentation results. It provides an additional layer of confidence in the

https://doi.org/10.5194/egusphere-2023-1819
Preprint. Discussion started: 14 August 2023

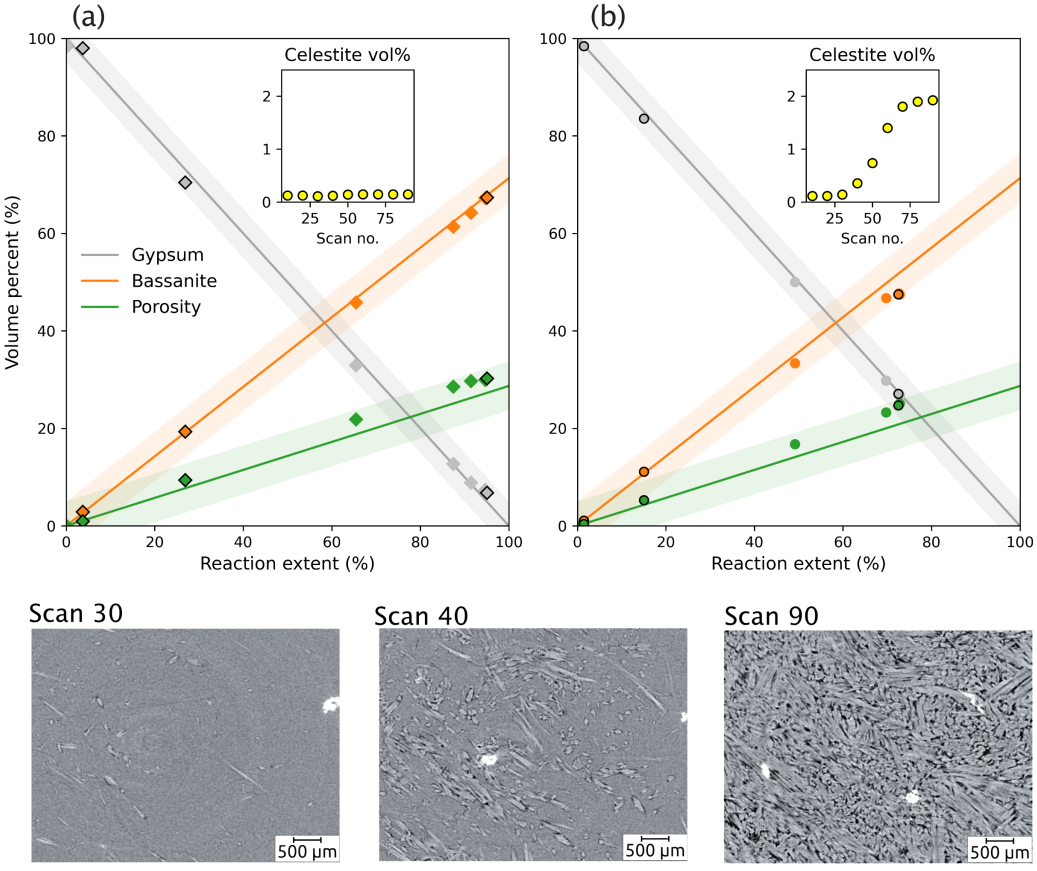

Figure 7. Quantitative analysis of phase volume changes during the gypsum dehydration experiment, as determined by segmenting the same volume VA19 time step 40 using two distinct segmentation methods. The new workflow developed in this study (a), which leverages a random forest classifier to label input data for a deep learning model, yields significantly improved accuracy in phase volume measurements relative to conventional thresholding segmentation – including "despeckle" and "non-local means" with sigma = 5, smoothing = 1 (b). The inset graphs show volume measurements for the celestite phase (yellow), which is a non-reacting phase during dehydration. The bottom images show slices of the sample at different stages; in the graph they are represented by the data points with a black outline

accuracy of our measurements, ensuring that the segmentation method effectively captures the phase evolution within the rock samples.

4.2 General applicability of the proposed workflow

355 The versatility of the presented workflow extends beyond the study of the gypsum dehydration process. By leveraging 4D μCT imaging and integrating chemical knowledge, our approach has potential for investigating other fluid-rock interaction processes, enabling precise quantification of mineralogical changes, and providing valuable insights into various geological

https://doi.org/10.5194/egusphere-2023-1819
Preprint. Discussion started: 14 August 2023

phenomena. For example, our approach can be directly applied to the investigation of the $KBr-KCl$ solid-solid replacement, which serves as an analogue for studying the dolomitization mechanism and other solvent-mediated reactions, resulting in the creation of porosity (Beaudoin et al., 2018). Similarly, the method has potential in fluid-rock interaction reactions relevant in the geoenergy field: our methodology can contribute to the analysis of carbonation reactions within ultramafic rocks, where carbon dioxide (CO_2) reacts with minerals to form carbonate minerals (Beinlich et al., 2020; Snaebjörnsdóttir et al., 2020), thus gain valuable insights into the mineralogical changes associated with carbon dioxide (CO_2) sequestration, contributing to the development of efficient carbon capture strategies. Additionally, our method is applicable to studying metasomatic and alteration processes related to hydrothermal fluids, shedding light on transformations occurring in geothermal reservoirs (Heap et al., 2020).

The proposed approach enables us to quantify geological processes at the grain scale, integrating with data from other sources and a priori chemical knowledge. This synergy between advanced imaging techniques and chemical understanding can bring about a new level of precision in our comprehension of complex geological processes. The ability to capture and analyse the temporal evolution of mineral phases with high spatial resolution provides us with a detailed understanding of the dynamic behaviour of geological systems. This enhanced level of insight allows us to unravel the intricate mechanisms governing rock deformation, metamorphic processes, and fluid-rock interactions.

4.3 Future horizons of deep learning segmentation for image analysis in geosciences

The success of our deep learning methods in the task of segmenting complex 4D data can represent a versatile approach that can find use in many image analysis tasks of geomaterials. By providing a reusable and adaptable workflow, we open the door to collaborations and innovations within the scientific community.

In future iterations of our method aims to expand its capabilities and applications. A direction to explore is the integration of deep learning convolutional neural networks with transfer learning and reinforcement learning techniques. Transfer learning can leverage pre-trained models to reduce computational cost and improve generalisation ability (Kim et al., 2022), while reinforcement learning might provide dynamic and adaptive strategies for data acquisition and reconstruction (Le et al., 2022). For our case study, by using the chemical theoretical molar reaction as a guiding principle, we can train the segmentation algorithm to identify and accurately outline the volumes of different mineral phases at various stages of the dehydration process. This adaptive learning process, driven by the theoretical molar reaction, could maintain high accuracy and robustness of the segmentation algorithm throughout the dehydration process. In addition to this promising integration of techniques, two key areas of potential advancement lie in the development of unsupervised segmentation approaches and the use of time as a parameter to learn from. Unsupervised learning can dramatically reduce the time and effort required for data annotation, accelerating analysis, and enabling the exploration of larger datasets (Mahdaviara et al., 2023). While data from before and after a scan in a time series may provide extra information that can be leveraged to better segment complex datasets. 4D data pose a unique challenge and opportunity for these unsupervised methods, as leveraging temporal information can significantly improve the quality and consistency of the segmentation.

https://doi.org/10.5194/egusphere-2023-1819
Preprint. Discussion started: 14 August 2023

5 Conclusions

In this work, we have demonstrated the potential of deep learning methods in the segmentation of 4D synchrotron X-ray tomographic images, particularly in the context of metamorphic rock transformations. We successfully overcame the inherent challenge of accurately segmenting all mineral phases and the pore network in an operando dataset, consisting of around 50 tomograms for each experimental setting, by using a robust and efficient deep learning-based workflow.

Our deep learning algorithm, trained on just 13 representative slices, generated a reliable segmentation, substantiating the versatility and power of such approaches. Conversely to the conventional external calibration techniques, we achieved validation of the segmentation accuracy by employing the metamorphic reactions themselves as an internal standard. We found the errors between the theoretical and segmented volumes from our time-series experiments to be consistently within the 2% confidence intervals of the theoretical curves. This facilitates extracting quantitative information, such as mineral growth rate and pore size variations, from segmented CT images during a reaction. The implementation of a 2D U-net architecture for segmentation and the utilisation of Random Forest-obtained labelled data as input demonstrated how machine learning can efficiently process large datasets and provide robust results even under challenging conditions. Coupled with the advantage of very short run times, our algorithm demonstrates great potential for practical application in similar studies.

In conclusion, our study underscores the transformative potential of deep learning in the realm of image analysis for geomaterials. The robustness, accuracy, and efficiency of our algorithm, coupled with its reusability, highlight how such methods can significantly advance research in this field. We anticipate that our approach will serve as a catalyst for further research, empowering scientists to make accurate predictions about microstructural changes under various stress conditions and contributing to a deeper understanding of tectono-metamorphic processes. We encourage other researchers to adopt and develop the workflow we introduced here, fostering an environment of shared learning and collaboration within the scientific community.

Code availability. Data analysis and plots were created using the Matplotlib library for the Python language (https://matplotlib.org/stable/index.html); the script for recreating the figures together with the input data are available at: https://doi.org/10.7488/ds/7493.

Data availability. The images and deep learning models for this paper were generated using Dragonfly software, Version 2020.2 for Windows. Object Research Systems (ORS) Inc, Montreal, Canada, 2020; software available at http://www.theobjects.com/dragonfly. The deep leaning model and data set used in this work are available at:

- Deep learning model: https://doi.org/10.7488/ds/7493
- VA17: https://doi.org/10.16907/8ca0995b-d09b-46a7-945d-a996a70bf70b
- VA19: https://doi.org/10.16907/a97b5230-7a16-4fdf-92f6-1ed800e45e37

https://doi.org/10.5194/egusphere-2023-1819
Preprint. Discussion started: 14 August 2023

Appendix A: Manual Segmentation

Manual Segmentation is performed using the Dragony software. The different features of interest are identified by the human eye and we define the intensity range of grey value according to the specific material phase.

Appendix B: Random Forest Segmentation

Random Forest pixel classification is performed using the Dragonfly software. The pixels pertaining to the different phases (gypsum, bassanite, pore, celestite) visible in the sample are identified and painted using the Brush tool in the software. We manually classified phases over a small number of slices (i.e., thirteen slices) and then used these data as input dataset into the Random Forest classifier. A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control overfitting. In our case, the algorithm is a pixel-based segmentation computed here using local features based on local intensity, edges and textures at different scales. The pixels of the mask are used to train a random-forest classifier from scikit-learn (Pedregosa et al., 2011). Intensity, gradient intensity and local structure are computed at different scales thanks to Gaussian blurring.

Appendix C: Evaluation Metrics Parameters

To help evaluate Deep Learning segmentation quality, we use a set of different evaluation metrics for comparing the neural network models trained with different ground truth data. All presented metrics are based on the computation of a confusion matrix for the segmentation task. The confusion matrix is built on the so-called “basic cardinalities” which can be calculated within the Dragonfly software. Basic cardinalities include the number of true positive (TP), false positive (FP), true negative (TN), and false negative (FN) predictions. For a full mathematical description of the cardinalities we refer to Taha & Hanbury (2015) and to Müller et al. (2022). For all metrics shown here, except Cohen’s Kappa, the value ranges from zero (worst) to one (best).

C1 Recall, Specificity, and Precision

Recall, also known as sensitivity or true positive rate (TPR), focuses on the true positive detection capabilities. Specificity, instead, evaluates the ability for correctly identifying true negative classes, thus, it is also known as true negative rate. Another related measure is Precision, also called positive predictive value (PPV), which is not commonly used in validation of tomographic images, but it is used to calculate the F-measure (see below). These three metrics are calculated as:

https://doi.org/10.5194/egusphere-2023-1819
Preprint. Discussion started: 14 August 2023

$$Recall = \frac{TP}{TP + FN} \tag{C1}$$

(C2)

$$Specificity = \frac{TN}{TN + FP} \tag{C3}$$

(C4)

$$Precision = \frac{TP}{TP + FP} \tag{C5}$$

C2 Accuracy

Accuracy is one the most known evaluation metric in statistics (Müller et al., 2022). It is defined as the number of correct predictions, consisting of true positives and true negatives, compared to the total number of predictions. However, many recent works (see Taha & Hanbury, 2015; Müller et al., 2022, for a complete review) have discouraged the use of accuracy in image analysis, particularly in multi-class segmentation where class imbalance is highly common: because of the true negative inclusion, the accuracy metric will always result in an anomalous high scoring (Müller et al., 2022). This can be clearly seen in Figure A1, where the score for the Accuracy metric is high also for those models which do not perform well if taking into account other metrics. Accuracy is calculated as:

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \tag{C6}$$

C3 F-measure based metrics

F-measure, also known as F-score, metrics are among the most widely used evaluation performance metrics for computer vision and image analysis (Taha & Hanbury, 2015; Müller et al., 2021, 2022; Allen et al., 2022). It is calculated from Recall and Precision of a prediction, by which it scores the overlap between predicted segmentation and ground truth. Including the precision metric, F-measure penalises false positives, which can be common features in multi-class datasets – such as those derived from X-ray μCT. There are two metrics based on the F-measure: Dice Coefficient, also called F1 or Sørensen-Dice index, and the Intersection-over-Union (IoU), also known as Jaccard index or Jaccard similarity coefficient. The Dice coefficient is defined as the harmonic mean between sensitivity and precision and is calculated as:

$$DICE = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \tag{C7}$$

The IoU, instead, is defined as:

$$IoU = \frac{TP}{TP + FP + FN} \tag{C8}$$

https://doi.org/10.5194/egusphere-2023-1819
Preprint. Discussion started: 14 August 2023

We can also define DICE as:

$$DICE = \frac{2 \times IoU}{1 \times IoU} \tag{C9}$$

C4 Area under the Receiver Operating Characteristic

The Receiver Operating Characteristic (ROC), is a line plot of the diagnostic ability of a classifier by visualising its performance with different discrimination thresholds (Taha & Hanbury, 2015; Müller et al., 2022). The performance is assessed through the true positive rate against the false negative rate. We can use the Area under the Receiver Operating Characteristic (AUC) as a single-value evaluation performance metric for the validation of image classifiers (Müller et al., 2022). The following AUC formula is determined as the area of the trapezoid defined by the ROC plot (see Müller et al. (2022) for a full formulation):

$$AUC = 1 - \frac{1}{2}\left(\frac{FP}{FP+TN} + \frac{FN}{FN+TP}\right) \tag{C10}$$

It needs to be noted that an AUC value of 0.5 is indicative of a random classifier.

C5 Volumetric similarity

As the name suggests, Volumetric Similarity (VS) is a measure that considers the volume of the segmented classes to indicate similarity. Here we use the definition reported in Taha & Hanbury (2015), namely the absolute volume difference divided by the sum of the compared volumes. Taha & Hanbury (2015) define the VS as 1-VD, where VD is the volumetric distance:

$$VolumetricSimilarity = 1 - \frac{|FN - FP|}{2 \cdot TP + FP + FN} \tag{C11}$$

C6 Cohen's Kappa

This metric is defined as a change-corrected measure of agreement between ground truth and predicted classification (Taha & Hanbury, 2015; Müller et al., 2022). Differently for previous metrics, Cohen's Kappa (KAPPA) ranges from -1 (worst) and +1 (best); a KAPPA close to 0 indicates a random classifier. The KAPPA evaluation metric is calculated as follows:

$$f_c = \frac{(TN+FN)(TN+FP)+(FP+TP)(FN+TP)}{TP+TN+FN+FP} \tag{C12}$$

(C13)

$$Kappa = \frac{(TP+TN)-fc}{(TN+TN+FN+FP)-fc} \tag{C14}$$

In the main text the Dice coefficient (Fig. 4 and Table 1) is used to evaluate and compare the segmentation resulting from the neural network trained using ground truth data derived from (i) Histogram segmentation (Models A, B, C), (ii) Histogram

https://doi.org/10.5194/egusphere-2023-1819
Preprint. Discussion started: 14 August 2023

segmentation with data augmentation (Models D and E), and finally (iii) a Random Forest Classifier. A complete description of all calculated metrics can be found in Figure A1 and in Table A1. Both the figure and the table report the calculated values for the different phases (Gypsum, Bassanite, Pores, and Celestite) and the average over the segmented volume, for the reference volume VA19-040 (736x800x400 voxels). It can be noted how the introduction of data augmentation benefits the segmentation of most phases with respect to almost all metrics (particularly for Model E). However, only Model RF (trained with a Random Forest ground truth) includes the Celestite phase (yellow in the graphs) in addition to the overall best performance in all most metrics.

Author contributions. Roberto Emanuele Rizzo: Conceptualization, Methodology, Investigation, Formal analysis, Writing - Original Draft. Damien Freitas: Conceptualisation, Investigation, Formal analysis, Data curation, Writing- Original draft. James Gilgannon: Methodology, Investigation, Formal analysis, Writing- Original draft. Sohan Seth: Validation, Resources, Writing - Review and Editing. Ian B. Butler: Investigation, Conceptualisation, Validation, Funding acquisition, Writing- Reviewing and Editing. Gina McGill: Investigation, Data curation, Writing- Reviewing and Editing. Florian Fusseis: Investigation, Conceptualisation, Supervision, Funding acquisition, Writing- Reviewing and Editing.

Competing interests. Dr. Florian Fusseis is a member of the editorial board of Solid Earth. The other authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements. This research was funded through NERC standard grant NE/T001615/1. We would like to thank Federica Marone and Christian Schlepütz at TOMCAT beamline at PSI for the fantastic support and assistance during beamtime. Oliver Plümper, Hamed Amiri, and Alireza Chogani from Utrecht University for invaluable help during the long days at the beamtime. Finally, we are very grateful to John Wheeler for discussions during the drafting of this paper. We are sincerely grateful to Object Research Systems (ORS) Inc (Montreal, Canada) for granting us a free-of-charge Non-Commercial licence of their software Dragonfly.

https://doi.org/10.5194/egusphere-2023-1819
Preprint. Discussion started: 14 August 2023

References

Adams, J. E. (2009). Quantitative computed tomography. European journal of radiology, 71(3), 415-424.

Allen, E., Lim, L. Y., Xiao, X., Liu, A., Toney, M. F., Cabana, J., & Nelson Weker, J. (2022). Spatial Quantification of Microstructural Degradation during Fast Charge in 18650 Lithium-Ion Batteries through Operando X-ray Microtomography and Euclidean Distance Mapping. ACS Applied Energy Materials, 5(10), 12798-12808.

Andrew, M. (2018). A quantified study of segmentation techniques on synthetic geological XRM and FIB-SEM images. Computational Geosciences, 22(6), 1503-1512.

Badran, A., Marshall, D., Legault, Z., Makovetsky, R., Provencher, B., Piché, N., & Marsh, M. (2020). Automated segmentation of computed tomography images of fiber-reinforced composites by deep learning. Journal of Materials Science, 55(34), 16273-16289

Beaudoin, N., Hamilton, A., Koehn, D., Shipton, Z. K., & Kelka, U. (2018). Reaction-induced porosity fingering: replacement dynamic and porosity evolution in the KBr-KCl system. Geochimica et Cosmochimica Acta, 232, 163-180.

Beinlich, A., Plümper, O., Boter, E., Müller, I.A., Kourim, F., Ziegler, M., Harigane, Y., Lafay, R., Kelemen, P.B. and Oman Drilling Project Science Team (2020). Ultramafic rock carbonation: Constraints from listvenite core BT1B, Oman Drilling Project. Journal of Geophysical Research: Solid Earth, 125(6), p.e2019JB019060.

Bizhani, M., Ardakani, O.H. & Little, E. Reconstructing high fidelity digital rock images using deep convolutional neural networks. Sci Rep 12, 4264 (2022).

Butler, I. B., Fusseis, F., Cartwright-Taylor, A. & Flynn, M. (2020), 'Mjölnir: a miniature triaxial rock deformation apparatus for 4D synchrotron x-ray micro-tomography', Journal of Synchrotron Radiation 27, 1681-1687.

Cartwright-Taylor, A., Mangriotis, M.-D., Main, I. G., Butler, I. B., Fusseis, F., Ling, M., Andò, E., Curtis, A., Bell, A. F., Crippen, A., Rizzo, R. E., Marti, S., Leung, D. & Magdysyuk, O. V. (2022) 'Seismic events miss important kinematically governed grain scale mechanisms during shear failure of porous rock'. Nature Communications 13, 6169.

Da Wang, Y., Blunt, M. J., Armstrong, R. T., & Mostaghimi, P. (2021). Deep learning in pore scale imaging and modeling. Earth-Science Reviews, 215, 103555.

Dice, L. R. (1945). Measures of the amount of ecologic association between species. Ecology, 26(3), 297-302.

Fusseis, F., Schrank, C., Liu, J., Karrech, A., Llana-Fúnez, S., Xiao, X., & Regenauer-Lieb, K. (2012). Pore formation during dehydration of a polycrystalline gypsum sample observed and quantified in a time-series synchrotron X-ray micro-tomography experiment. Solid Earth, 3(1), 71-86.

Fusseis, F., Schrank, C. Xiao, X. & De Carlo, F. (2014). The application of synchrotron radiation-based microtomography to (structural) geology, Journal of Structural Geology 65, 1-14.

Gilgannon, J., Freitas, D., Rizzo, R., Wheeler, J., Butler, I., Seth, S., Marone, F., Schlepütz, C., McGill, G., Watt, I. & Plümper, O., 2023. A non-hydrostatic stress state forms fabrics during metamorphic reactions (No. EGU23-6483). Copernicus Meetings.

Heap, M. J., Gravley, D. M., Kennedy, B. M., Gilg, H. A., Bertolett, E., & Barker, S. L. (2020). Quantifying the role of hydrothermal alteration in creating geothermal and epithermal mineral resources: The Ohakuri ignimbrite (Taupō Volcanic Zone, New Zealand). Journal of Volcanology and Geothermal Research, 390, 106703.

Karimpouli, S., & Tahmasebi, P. (2019). Segmentation of digital rock images using deep convolutional autoencoder networks. Computers & geosciences, 126, 142-150.

https://doi.org/10.5194/egusphere-2023-1819
Preprint. Discussion started: 14 August 2023

Kim, H. E., Cosa-Linan, A., Santhanam, N., Jannesari, M., Maros, M. E., & Ganslandt, T. (2022). Transfer learning for medical image classification: a literature review. BMC medical imaging, 22(1), 69.

Kruth, J. P., Bartscher, M., Carmignato, S., Schmitt, R., De Chiffre, L., & Weckenmann, A. (2011). Computed tomography for dimensional metrology. CIRP annals, 60(2), 821-842.

Le, N., Rathour, V. S., Yamazaki, K., Luu, K., & Savvides, M. (2022). Deep reinforcement learning in computer vision: a comprehensive survey. Artificial Intelligence Review, 1-87.

Lee, D., Karadimitriou, N., Ruf, M., & Steeb, H. (2022). Detecting micro fractures: a comprehensive comparison of conventional and machine-learning-based segmentation methods. Solid Earth, 13(9), 1475-1494.

Mahdaviara, M., Sharifi, M., & Rafiei, Y. (2023). PoreSeg: An Unsupervised and Interactive-based Framework for Automatic Segmentation of X-ray Tomography of Porous Materials. Advances in Water Resources, 104495.

Marti, S., Fusseis, F., Butler, I.B., Schlepütz, C., Marone, F., Gilgannon, J., Kilian, R. & Yang, Y., (2021). Time-resolved grain-scale 3D imaging of hydrofracturing in halite layers induced by gypsum dehydration and pore fluid pressure buildup. Earth and Planetary Science Letters, 554, p.116679.

Müller, D., Soto-Rey, I., & Kramer, F. (2022). Towards a guideline for evaluation metrics in medical image segmentation. BMC Research Notes, 15(1), 1-8.

Müller, S., Sauter, C., Shunmugasundaram, R., Wenzler, N., De Andrade, V., De Carlo, F., Konukoglu, E. & Wood, V. (2021). Deep learning-based segmentation of lithium-ion battery microstructures enhanced by artificially generated electrodes. Nature communications, 12(1), pp.1-12.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. & Vanderplas, J., (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, pp.2825-2830

Phan, J., Ruspini, L. C., & Lindseth, F. (2021). Automatic segmentation tool for 3D digital rocks by deep learning. Scientific Reports, 11(1), 1-15.

Phillips, T., Bultreys, T., Bisdom, K., Kampman, N., Van Offenwert, S., Mascini, A., Cnudde, V. & Busch, A. (2021). A Systematic Investigation Into the Control of Roughness on the Flow Properties of 3D-Printed Fractures. Water Resources Research, 57(4), pp. ewrcr-

Rezaee, M., Mahdianpari, M., Zhang, Y., & Salehi, B. (2018). Deep convolutional neural network for complex wetland classification using optical remote sensing imagery. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 11(9), 3030-3039.

Reinhardt, M., Jacob, A., Sadeghnejad, S., Cappuccio, F., Arnold, P., Frank, S., Enzmann, F. & Kersten, M., (2022). Benchmarking conventional and machine learning segmentation techniques for digital rock physics analysis of fractured rocks. Environmental Earth Sciences, 81(3), p.71.

Renard, F., Guedria, S., Palma, N. D., & Vuillerme, N. (2020). Variability and reproducibility in deep learning for medical image segmentation. Scientific Reports, 10(1), 1-16.

Ronneberger, O.; Fischer, P.; & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In Lecture Notes in Computer Science, Springer International Publishing, Vol. 9351, pp 234-241.

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. Journal of big data, 6(1), 1-48

Snaebjörnsdóttir, S. Ó., Sigfússon, B., Marieni, C., Goldberg, D., Gislason, S. R., & Oelkers, E. H. (2020). Carbon dioxide storage through mineral carbonation. Nature Reviews Earth & Environment, 1(2), 90-102.

Sokac, M., Budak, I., Katic, M., Jakovljevic, Z., Santosi, Z., & Vukelic, D. (2020). Improved surface extraction of multi-material components for single-source industrial X-ray computed tomography. Measurement, 153, 107438

https://doi.org/10.5194/egusphere-2023-1819
Preprint. Discussion started: 14 August 2023

Taha, A. A., & Hanbury, A. (2015). Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC medical imaging, 15(1), 1-28.

Torralba, M., Jiménez, R., Yagüe-Fabra, J. A., Ontiveros, S., & Tosello, G. (2018). Comparison of surface extraction techniques performance in computed tomography for 3D complex micro-geometry dimensional measurements. The International Journal of Advanced Manufacturing Technology, 97, 441-453.

Withers, P.J., Bouman, C., Carmignato, S., Cnudde, V., Grimaldi, D., Hagen, C.K., Maire, E., Manley, M., Du Plessis, A. & Stock, S.R., (2021). X-ray computed tomography. Nature Reviews Methods Primers, 1(1), p.18.

Zeiler, M. D. (2012). Adadelta: an adaptive learning rate method. arXiv preprint arXiv:1212.5701.

https://doi.org/10.5194/egusphere-2023-1819
Preprint. Discussion started: 14 August 2023

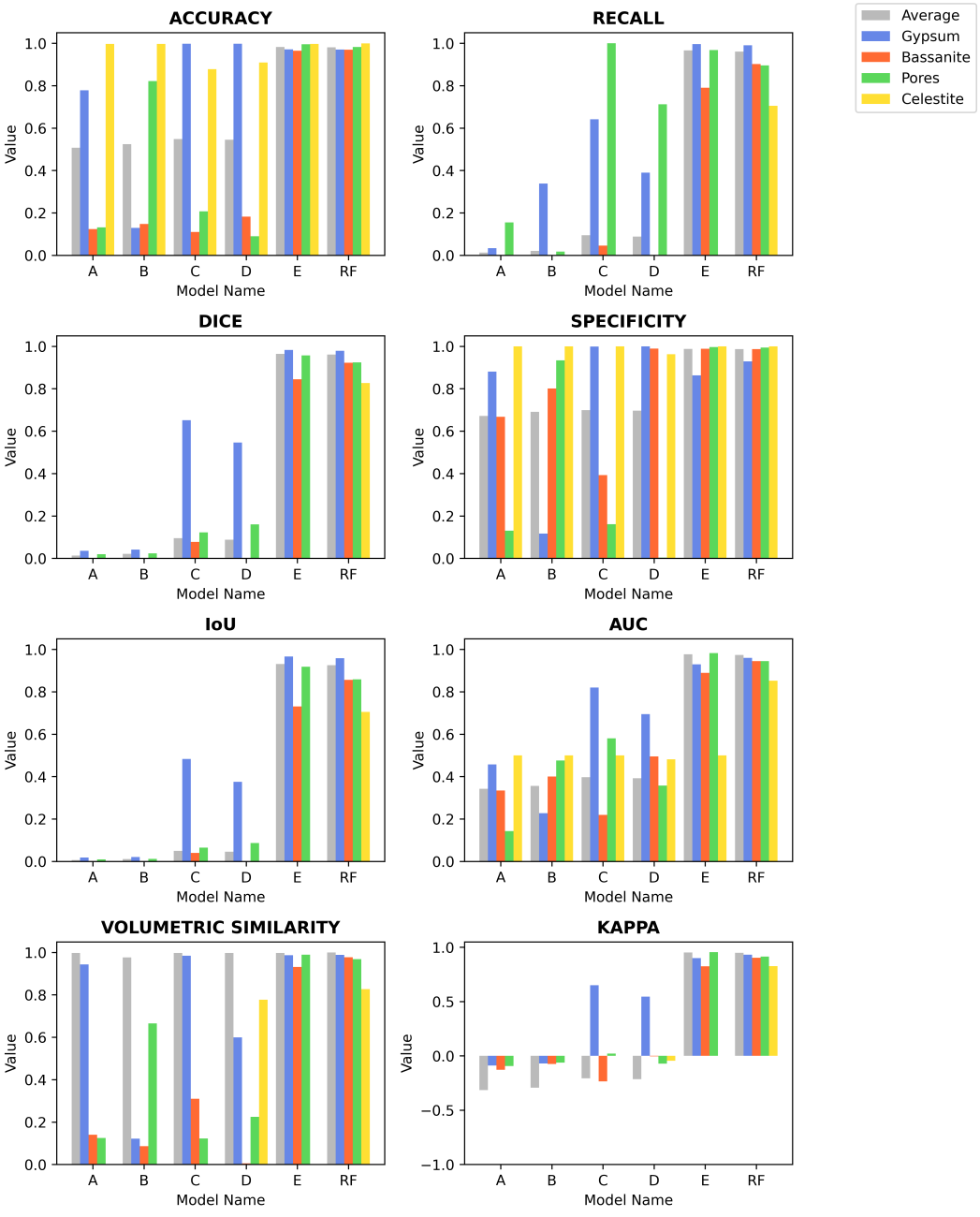

Figure A1. Eight common evaluation metrics calculated for the different segmentation models: Accuracy, Recall, Dice Coefficient (DICE), Specificity, Intersection-over-Union (IoU), Area under the Receiver Operating Characteristic (AUC), Volumetric Similarity, and Cohen's Kappa (Kappa) are evaluated for the average volume and for each of the phases present in the analysed sample volume. Please refer to the main text for details regarding each trained model.