Reviewer #3 – Luke Griffith

Dear editor and authors,

Summary: This paper presents a workflow for multi-class segmentation of X-ray computed tomography images of rock during in situ testing. The authors aim to address the reproducibility issues commonly associated with this process through employing a deep learning approach, trained on the results of a feature-based random forest classification. The results of the U-net model are compared to a global thresholding method. They innovatively use mass balance to provide an external measure of the relative volume of constituents, allowing for a direct comparison with the segmentation results.

The paper is well-written and the background and motivation are clearly outlined. I have some comments on the methodology, for which some choices need to be explained in greater detail and potentially modified where appropriate to compare the models and evaluate their performance in a more robust way. I found the external verification of the multi-class segmentation to be a clear strong point of the paper, and I think this data could be further used as *a priori* information to guide the models during training and provide robust segmentation (but perhaps this is beyond the scope of this study).

I have provided comments below, and comments in the attached PDF on specific parts of the text.

I recommend publication after minor revisions, most importantly to address the questions on the segmentation methodology.

We are sincerely grateful to Dr. Luke Griffith for his thorough and detailed review of our manuscript. We have addressed the comments made here below, reporting also those made on the PDF version of the manuscript.

Comments:

- The distinction between pixel or feature-based methods and convolutional methods could perhaps be made clearer earlier on – the assumption is that convolutional methods work better because they are able to better capture spatial information but I was expecting this point to have more weight in the introduction.

  We thank the reviewer for point out that these concepts were not clearly described in the Introduction section. We have integrated the text as follows:
  "In addition, some deep learning algorithms still rely on adaptive filtering and global thresholding operations (Phan et al., 2021). This reliance on the grayscale value can hinder the effectiveness of such

algorithms, regardless of their complexity. This limitation becomes most apparent in data containing low contrast phases, where filtering processes to reduce noise or enhance feature visibility may alter or eliminate critical intensity variations necessary for accurate phase differentiation and segmentation. In contrast, convolutional methods, grounded in machine learning, advance beyond these constraints by integrating spatial and morphological information. This integration allows for a more robust and accurate segmentation, especially vital in μCT datasets, where spatial relationships and contextual nuances are key to discerning accurate interpretations."

- Was there any attempt to use histogram matching and/or calibration of greyscale to values of known materials (e.g. for pieces of steel that are not expected to change between scans)? This can work quite well to allow for direct comparison of time-lapse images and quantitative analysis.
  We agree that using a phantom of a known material can help calibrating the grayscale of the sample. We have discussed this method in the manuscript, but for the dataset shown in the manuscript we did not use a phantom. However, in the upcoming beamtime we will test the method.

- The ML approaches are compared to simple histogram (global?) thresholding – perhaps even adaptive/local thresholding or watershed segmentation would perform much better and might be a fairer comparison.

  We agree the adaptative/local thresholding shows better performances that histogram thresholding. However, we deliberately used histogram thresholding as comparison since the this is still one of the most commonly used method for in image segmentation.

- How did you label the training data? Do you label all pixels within an image or a subset of pixels for which you are confident in their labelling (e.g. only pixels well within a grain)?

  As reported in Appendix B of the manuscript we labelled the training dataset using the Dragonfly software. We labelled pixels pertaining to the different phases (gypsum, bassanite, pore, celestite) using the Brush tool in the software only within the mineral grains or pores.

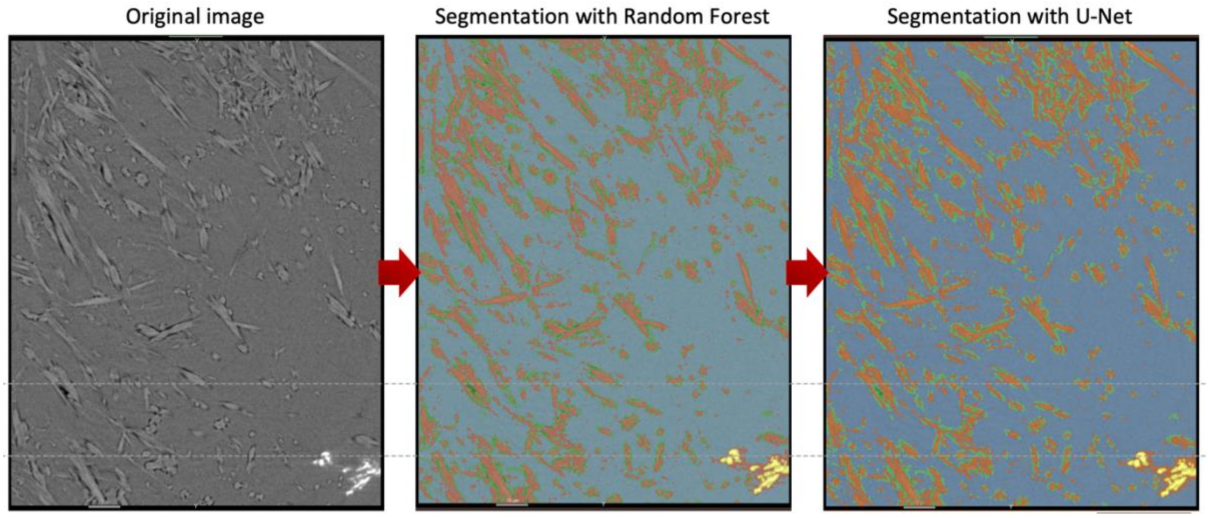
- How was the validation set chosen? Equally across all images from each time step? I would be interested to see how the model performs if the validation set contained all images of a given time step (therefore not present in the training set). As it is argued that the DL method is more robust and consistent, doing the test-train split in this way could better illustrate this point.

For our latest model (i.e. that trained using RF segmented data as input) we defined the "validation set " randomly setting aside 20% of the input data. In our first attempts to train a DL model, we tended to feed as ground truth, training and validation sets large volumes of data. This, however, induced the model to overfit. Thus the choice of training the data over small datasets integrated with Data Augmentation.

- How does the RF model alone compare to the U-net model? For example, compared to the manually labelled images.

  The RF and DL with U-net architecture models produce equally comparable (and good) results for the dataset used for training (see below figure). However the accuracy of the RF segmentation tends to degrade moving towards the extremes of the time-series data (i.e., early stage of the dehydration and near-fully dehydrated samples) as well as when attempting to segment unseen data.

  Differently the DL model is consistent in the segmentation for different (unseen) XCT volumes and throughout the time-series.



- How were the features chosen for the RF model? How much better could the RF model become if more features are added and/or features with larger windows than 3x3? Could it rival the U-net model?

  To clarify the method used, we have now integrated more details on the RF features into the text of Appendix B:

  "In our study, the Random Forest classifier was employed with a set of predefined features: Morphological, Gaussian Multi-Scale, and

Neighbours. Each of these feature sets plays a distinct role in enhancing the classifier’s ability to accurately segment phases in the dataset.

– Morphological Features: These are used to analyse the shape and structure within the images, enabling the classifier to detect and distinguish different phases based on their morphological characteristics.
– Gaussian Multi-Scale Features: These features involve applying Gaussian filters at multiple scales, aiding in smoothing the images and reducing noise. This multi-scale approach helps in capturing features at various levels of detail, contributing to more effective phase differentiation.
– Neighbours Features: This set focuses on the local neighbourhood of each pixel, capturing the texture and local contrast, which is essential for identifying subtle boundaries between phases.

All these features are used together in the Random Forest classifier, each contributing to the overall classification task. The classifier does not operate on a voting system between these feature sets; rather, it integrates the information provided by all of them to decide for each voxel in the image. This integrated approach enables a more nuanced and accurate classification compared to using any single feature set on its own, and significantly improves the process over manual thresholding methods”

Regarding changing/modifying the features and their sizes. This is a very interesting point which we have not yet tested, but which is important to explore. Instead regarding the ability of performing more accurate segmentation compared to the U-net, we still believe that the main advantage of using a DL model stays in its ability to generalise. Random Forest methods (such as the Weka implementation in ImageJ; Arganda-Carreras et al., 2017 Bioinformatics, doi:10.1093/bioinformatics/btx180) are fully able to segment complex dataset but require constant ‘re-training’ when dealing with unseen datasets.

- Perhaps I misunderstand, but it is a bit confusing to use “ground truth” and *validation data* interchangeably (is this indeed the case?). For example, are the “ground truths” used for comparison in Table 1 the same for each method? I think, ideally, a ground truth should be the best possible segmentation (e.g., one that was labelled manually or using whichever is deemed the best result from all these models) and it should be the same reference for all models.
It can be misleading to evaluate the models on how well they perform on their validation data, because the quality of the validation data varies depending on the method used to label it (i.e., thresholding or

RF). At the very least, this should be made more clear by dropping the “ground truth” terminology. Ideally, the models should be re-evaluated against the same “best-case” segmentation.

We agree that there might be some confusion. When we refer to “ground truth” we refer to the labelling of data done my us as obtained using different segmentation strategies (i.e., Histogram thresholding or Random Forest Classification). Essentially, ground truth acts as the standard against which the model's predictions are compared. It would be not correct to compare a ground truth derived using random forest against the result of a DL model trained using Histogram Segmentation.

- Are you able to give a formal comparison of the methods based on the ground truth from the calculated phase volume changes? That seems to be like a good way to evaluate their performance.

  We have provided a formal comparison of the methods, both for the Deep Learning models trained using different input sets (Table 1 and Figure A1 in appendix), as well as of Deep Learning vs Histogram thresholding in Figure 7.

- More of an observation as it is perhaps beyond the scope of this work: it would be very interesting to see if the measured volumetric evolution could be used to constrain the models used for segmentation. This is what I was anticipating based on the title.

  This is indeed a very interesting topic. We think this can be incorporated into a “semi-supervised” deep-learning segmentation algorithm. We are indeed working towards including this into the deep-learning model.

- "U-net model are compared to a global thresholding method" - should read "U-net models trained on RF-generated labels are compared to U-net models trained on labels made using thresholding"

  thanks for clarifying

Specific Comments:

Line 52: Most, perhaps, but there still are many multi-label classification examples out there.

We agree, and we do mention few of them in the introduction. But the majority of the studies, particularly for porous media, fluid flow in porous media, and for solid state reaction – where µCT is largely used – focus on void and solid material classifications.

Line 53: I don't quite understand this sentence and how it relates to "this". And does the limitation refer to images, or the limited information within a single grayscale pixel? It seems to me that images contain a significant amount of information - hence the proposed use of CNNs?

Also Reviewer 1 raised that the wording in this paragraph was unclear. We now added the following sentence: "This is most clearly seen in data that contain low contrast phases, for which filtering processes to reduce noise or enhance feature visibility may modify or remove variations in intensity that are critical for accurate phase differentiation and segmentation"

Line 54: Is this referring to grayscale changes due to changes in the imaged object or due to instability of the imaging equipment itself? If the latter, I understand why simple comparison between images will cause issues, but I don't see why thresholds would need to change, otherwise. This could use a little more explanation.

In our manuscript, we refer to the changes in the imaged object over time, not to any instability in the imaging equipment.

As our system evolves, the grayscale values within the image volumes also change dynamically. This variation is highlighted in Figure 2, where it becomes evident that a fixed threshold value selected at one time frame becomes inadequate for accurately segmenting a certain phase at a later time. It's not just the variation in pixel count that we observe; there's also a shift in the average grayscale values of the pixels over time.

This means that the threshold that may be appropriate for one time frame may no longer be suitable as the system evolves, leading to inaccuracies in phase segmentation. Our study emphasises the challenges of using fixed thresholding in evolving systems, as the grayscale properties of each phase can change significantly over time, necessitating a more adaptable segmentation approach.

Line 57: For me, this is the main point. Thresholding is pixel based, and DL can account for a wider context

We agree. But we would like to emphasise that in our work was actually the ML model (the Random Forest Classifier) that allows detection and incorporation of morphological feature in the segmentation model.

Figure 1: “"manually labelling"?” instead of “Labelling”.

Modified as suggested

Line 137: How was this 20% chosen?

The 20% of data for validation is chosen randomly from the input data. We have now clarified this in the text as follows: “For all the tested strategies, we randomly choose a twenty percent of the segmented data to serve as a “validation set” which is otherwise not used during training.”

Line 185: Can be confusing to use "features" outside of the ML context.

Here we actually mean the textural/fabric features of the rock.

Line 191: how does this relate to the grain size, I imagine it is quite small?

Initially, the grainsize of the bassanite is indeed quite small. However, using a filter size of 3x3 minimises the possibility of smoothing (and therefore eliminating) pixels belonging to the bassanite phase particularly during the early stages of the dehydration experiment.

Line 216: Wording seems strange to me here

To us the meaning of this sentence is quite straightforward: If a DICE score is close to 0 it means that the segmentation does not match the ground truth. Otherwise, the closer the DICE score is to 1 the better the match between segmentation and ground truth is.

Line 271 – 272: Perhaps need to be more specific here as this has been done for porosity (e.g. Iassonov, 2009), and to calibrate fluid saturation measurements in CT imaging.

Thanks for pointing out the reference to the work of Iassonov. We have now deleted the sentence.

Line 293: Why is 5% chosen? Could you mean 95%? Also, how is it calculated?

Thanks for flagging this. Here we intend that the data points fall within a 5% “error bound”, and not “confidence interval” as erroneously stated in the manuscript. Now fixed. We choose 5% error bound as this is a standard value in statistics. The error is measured as distance between the data point and the theoretical curve.

Line 320: To some extent, one might use this as a first pass but I think watershed segmentation is more typical

We have now added watershed segmentation to integrate the text of the manuscript.

Line 328: Is this shown here?

Yes, we have shown how the combination of RF and DL can produce a segmentation model that is accurate and generalised over a full time series.

Line 340 – 347: It is not clear to me how the partial volume effect is relevant here - are there crystals which are on the order of the voxel size? Or mixing of the materials (effective medium)? If so, this should be explained.

Often, primarily due to resolution limitations, placing a precise boundary between two or more phases is challenging. Typically, a gradual transition is observed from the grayscale values of one phase to those of another, making it difficult to derive accurate volumetric data.

Line 349: Unless I misunderstand, this is slightly misleading, as a "priori" suggests this information is somehow considered in the model, which I don't think it is - rather this information is used as validation?

As for definition, *a-priori* denotes "a knowledge which proceeds from theoretical deduction rather than from observation or experience." We therefore think that our word choice is correct.

Figure 7: Does it say which is which?

Later in the figure caption we say: "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). "

Line 377 – 380: This sentence sticks out a bit - can you give more information on why these methods? Transfer learning and reinforcement learning are quite large topics that may well add value, but it's unclear how/why they might be beneficial for this specific case. For example, transfer learning using models based on different time steps, or different materials? More concretely, why would RL be beneficial?

We have now integrated the text so that to clarify how we intend to use the two methods to integrate segmentation strategies. Here is the new paragraph:

"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). Specifically, transfer learning could be utilised to adapt models initially trained on datasets derived using imaging techniques which provide higher textural resolutions (such as Scanning Electron Microscope – SEM), thereby enhancing their ability to generalise to complex datasets with minimal retraining. Reinforcement learning could play a crucial role in optimising data acquisition and reconstruction processes. By applying reinforcement learning algorithms, we could develop systems that dynamically adjust acquisition parameters or reconstruction techniques based on real-time feedback, leading to more efficient and accurate image analysis. For instance, in time-evolving systems, reinforcement learning could be used to adaptively select optimal imaging parameters for each time step, based on the changes observed in the previous scans.”