Response to Reviewers' Comments on:

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

Response to Reviewer #1 – Anonymous

I believe the paper offers a significant contribution to scientific progress within the scope of the journal, and it will be of wide interest to the geoscience community, particularly in the field of 3D mineralogy and petrology. This paper introduces a new workflow for digital image processing, and aims to revolutionize current methodologies of XCT image processing in the field of in-situ time-evolving synchrotron XCT datasets, which are often very large in size and time-consuming to analyse due to the challenges in low contrast, noise, evolving mineral phases. I believe the proposed workflow is quite robust and valid, as the authors cross-check the accuracy of the proposed method against theoretical molar evolution of gypsum to bassanite during the reaction, and their measured values fall within 2% confidence intervals. In addition, they also provide a robust presentation and comparison with other more traditional methods of image segmentation, with and without data augmentation or machine-learning labelled ground truth data, and also provide suggestions to make the method even less time-consuming. Overall, I believe this method would be applicable in many fields and will greatly improve digital image analyses in time-evolving synchrotron XCT datasets or even standard XCT scans where mineral phases may overlap in intensity.

We sincerely thank Reviewer #1 for their thorough review and positive feedback on our manuscript. We are pleased to know our work is considered a valuable contribution to the Geosciences' community. We appreciate the constructive comments, which we have largely accepted and incorporated into the manuscript. Our responses are colour-coded in blue.

I would be interested in seeing how the accuracy of this workflow can be checked when there is no prior-knowledge of a reaction, or when there is no theoretical curve to check it against with. This might be the case for other geological processes, where for example different mechanisms play a role, and where the molar volumes of mineral phases are not so well known. How do the authors propose to check their workflow in those instances? It would be good to include this explanation in the discussion.

As explained in the Discussion section of our paper, a prevalent approach in domains like medical imaging and material science involves external calibration techniques. These techniques utilize phantoms—objects with predefined dimensions and/or compositions—as reference standards.

Phantoms function as external benchmarks, enabling the evaluation of segmentation method accuracy through imaging and segmentation. Assessing the segmented volumes or dimensions of these phantoms against their actual known values provides insights into the accuracy and dependability of segmentation techniques in μCT imaging.

In addition to these strategies, other methods that could be considered for benchmarking segmentation outputs include: (a) Comparative Studies: Validation can be achieved by conducting comparative studies with established, well-validated segmentation algorithms. Aligning new segmentation results with those from recognized methods enhances the credibility of the novel approach. (b) Synthetic or Simulated Data: Employing synthetic or simulated datasets, where the accurate segmentation is pre-determined, offers a robust validation tool. By applying the segmentation algorithm to these controlled environments, we can comprehensively evaluate its precision and reliability.

I found that the overall presentation quality could be improved as some concepts mentioned in the paper are too technical for a non – expert audience, and they need explanation.

In response to the detailed suggestions from Reviewer 1, we have incorporated explanations for technical terms and expressions into the manuscript. Further elaboration on these additions is available in our responses to the Detailed Comments section.

For instance, many readers will not be familiar with terms such as "supervised deep learning". What is a supervised deep learning method? How does it differ from an unsupervised one? The authors mentioned both concepts in the paper, yet they fail to explain what they are and how they differ. They also did not explain why they chose one rather than the other. It would be good to at least explain the difference between the 2 methods (since both are mentioned) and why the authors made that choice, so that the readers can better understand what may or may not work in other contexts where this workflow may help in the analysis.

As recommended, we have added a concise yet thorough explanation of Deep Learning in the context of image segmentation to the 'Introduction' section. We have also elaborated on the differences between Supervised and Unsupervised Deep Learning for image segmentation and explained our rationale for opting for Supervised segmentation in our study.

Furthermore, when possible, terminology belonging to machine-learning should be avoided, as this journal covers a great variety of topics, and while some readers may be familiar with terms such as "(hyper)-parameters", these may not always be clear to a non-expert reader. Why are they (hyper) parameters and not just parameters? I would suggest avoiding such technical terminology when possible, or if needed, then it needs some explanation. The authors explain what (hyper)-parameters they used, but it is not clear what (hyper)-parameters are.

While we understand that some readers might find unclear the use of terms such as "(hyper-)parameters", these are well-established and precise terms in the AI user community which have been used in the scientific literature for more that 10 years. We agree with the reviewers that we need to provide some explanation for the terms used, however we believe that is also important to use the correct terminology particularly if this is has recently introduced in the Geosciences community. A full explanation of the term (hyper-)parameter is provided for the comment in the Detailed Comments section.

Some concepts are introduced without explanation of if there is one, it is presented in different sections. I flagged in the commented text where I could: for instance, Random Forest is not cross-referenced with the section in the Appendix. I think introducing cross-referencing to these sections next to the concept would help non-expert readers (example: random forest, sec. 3.3, Appendix X).

We thank the reviewer for pointing this out. We have now updated the text so that it points to the Appendix.

The paper also contains some minor typos and small inaccuracies, like lack of introduction of acronyms.

We thank the reviewer for flagging these errors, we have now reviewed and corrected them, please see the "Detailed Comments" section here below where we reported the notes left by the reviewer on the PDF version of the manuscript.

Some of the figures could be improved and be bigger (not sure if it is the formatting of the generated pdf). It would be good to have an overall figure (with the grain of celestite) showing all the steps, including the post-processing cleaning up.

We have considered the suggestions made by Reviewer 1 and updated Figures 1 and 2 so that now they show larger image and bigger labels.

Overall, I think the paper is a great scientific contribution to the community. Providing minor revisions are made (specifically targeting the improvement of clarity for non-expert readers), I suggest that the paper is accepted for publication.

Once more we thank Reviewer 1 for the suggestions which have helped clarify some aspect of the manuscript.

Detailed Comments:

Line 28: why only micro and not nano? surely you can make this XCT so that it includes both nano and micro. Also the acronym has not been introduced before and needs explanation.

We prefer using the term μCT as x-ray microtomography is the standard term for the technique. XCT as an abbreviation for X-ray computed tomography is more consistent with medical CT.

Line 28 – 30: this sentence is too long, break it.

Now: "Time-resolved (4D) operando experiments in X-ray Computed Microtomography (μCT) scanners have emerged as a promising way of studying solid-state reactions offering unprecedented insight into mineral phases and volume changes. This 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."

Line 31: spatial resolutions.

Added as suggested

Line 53 – 55: Worth explaining that using any type of filtering may restrict even further the limited greyscale information available.

We have 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 55 – 56: ..also worth explaining to a non-expert audience that this is because the threshold may vary across different time steps.

Following Reviewer's suggestion we have modified the text as: "Moreover, as the grayscale image inputs vary over time in time-series datasets, the effectiveness of histogram thresholding diminishes. This is because the optimal threshold for one time step may not be applicable for others, leading to inconsistent or inaccurate segmentations."

Line 57: do you meant characteristics?

Modified as suggested.

Line 60: what is a supervised deep learning? and how is it different from an unsupervised one? worth explain why you chose this one in particular?

We have now added the following sentence to clarify the meaning of supervised and unsupervised deep learning: "Supervised deep learning is a type of machine learning where the model is trained on a labelled dataset. This means that each output produced by the model is paired with the correct output, enabling the model to learn by comparing its predictions to the actual outcomes. This contrasts with unsupervised deep learning, where the model attempts to identify patterns and relationships directly from the input data without labelled outcomes."

We opted for supervised deep learning primarily due to its accessibility, as it is supported by several commercial and freely available segmentation software (such as Avizo, Dragonfly, and ilastik), and its relative ease of implementation compared to unsupervised learning, which often requires high-performance computing clusters

Line 84: as in creation of more porosity.

That's correct. However, here in the text we are describing the chemical reaction, therefore we think is more accurate to talk about water and not porosity.

Line 90: any specific reason why you choose these 2 in particular?

The two samples were chosen because they represent "endmembers" experiments: with VA17 deformed under low differential pressure (with confining pressure > than differential), and VA19 under high differential pressure (and low confining).

To include the motivation we have modified the text as follows: "For the work presented in this paper, we focus on data from two specific experiments, chosen as 'end-member' scenarios for their distinct evolving mineral fabric during the dehydration process…"

Line 91: not to be picky but CT stands for computed tomography and you used this acronym before to indicate X-Ray microComputed Tomography, now you call synchrotron microtomography with the same acronym even though we lost the computed and we add synchrotron. perhaps use a different acronym? or just call it synchrotron XCT since you introduced XCT before?

We have answered a similar question earlier in the comments. For this work we think the use of µCT is more correct since the technique employed is indeed X-ray microtomography or µCT, with Synchrotron X-ray Microtomography as S-µCT.

Line 92: SLS not defined... well I know it means Swiss Light Source but you introduce acronyms without explaining them first!

Added to text in line 85.

Line 93: why 27KeV?

This is the peak energy value observed at TOMCAT beamline, after optimising the filters for imaging with the Mjolnir rig.

The filter white beam characteristic is defined by the synchrotron source and beamline parameters. It mainly depends on (non-extensive) the storage beams current and energy for the source and the type of insertion device (bending magnet, wiggler etc.) for the beamline. For these experiments, we filtered the lower energy range of the white beam for improving the imaging quality. The appropriate filtering quantity is determined as a function of the rig and sample materials (nature, thickness) to have an optimal number of photons on the detector. Filtering shifts the peak energy of the white beam and was here measured at 27 KeV. This energy is high enough to provide a good flux of photon able be transmitted through t the aluminium, pressure vessel and the sample.

Line 95: are these timings referring to those 2 samples or all experiments? unclear, do they vary between 150 and 314 minutes?

The time frame we are referring to here is indicative of the whole suite of experiments.

Line 98: can we make the figure bigger? especially the individual slices. I can see the changes but if the figures were a tad bigger that would be better

We have now increased the size of the image in the PDF file. We have also modified the figure to include the direction of the slices.

Line 100: why? needs explanation for non-expert readers earlier that this statement

To clarify the sentence we have now added: “… as the optimal threshold varies across different time steps”.

Line 117: reference formatting

Fixed

Figure 2 caption: “slice in the tomographic scan” instead of “individual image”. “slices” (and which direction) instead of “images”.

Modified as suggested

Line 126: what are network "hyper"-parameters. why can they not be called just parameters in this case?

In the context of machine learning and deep learning, 'parameters' and '(hyper-)parameters' refer to two distinct types of variables within a neural network model. Parameters are the internal elements of the model, such as weights and biases, that the model learns through training. These are adjusted automatically during the training process as the model learns from the data.

In contrast, (hyper-)parameters are the external settings of the model that are set prior to training and remain constant during the training process. These include choices such as learning rate, number of hidden layers, batch size, and epochs. Hyperparameters guide the learning process but are not learned from the data themselves. They are crucial for the model's architecture and learning process but require manual setting or optimization, often through experimentation.

Therefore, we refer to these settings as '(hyper-)parameters' to distinguish them from the learned 'parameters' of the network, as they play a different but equally crucial role in the development and performance of the model.

To clarify this we have modified the sentence in Line 126: “Choosing the best training neural network architecture and tuning the network (hyper-)parameters – i.e., those settings of the model that are set prior to training and remain constant during the training process – requires time and some knowledge of neural network architecture.”

Line 151: obtained through random forest?

This is a general behaviour of the Neural network architecture, independent from the kind of input data.

Line 153: "training"

Modified as suggested

Line 163 – 165: this sentence is unclear to me: you say for example i) histogram thresholding then ii) ..as provided by histogram thresholding.. what do you mean? it feels like a repetition, so please clarify the sentence. Also, the second part of the sentence is grammatically disconnected.

We have revised the sentence as follows: “To establish the ground truth set for initial image classification, we initially considered methods with differing informational depths. One such method is histogram thresholding, which relies on basic greyscale values and typically results in low-information-level outcomes. However, this approach alone proved inadequate for our purpose, as it often led to gradients with diffuse phase boundaries, underscoring the need for more sophisticated classification methods.”

Line 168: a visual quality check?

The check was done both visually but also using evaluation metrics (See figure in the appendix). We have now added a reference to the figure and the text in the appendix for clarity.

Line 171: perhaps it is me, but there are no values in brackets: better to say that they are indicated by different colours in the histogram curve?

For clarity we modified the text as follows: “The corresponding colour-coded threshold values for the four phases are shown in Figure 3b”

Line 176: I do not understand what this means. What does it mean that you octupled the input data in actual terms? Audience will less expertise on the matter will find hard to understand what this means. From this sentence alone, it is not clear to me what is data augmentation and how that improved the results of the data segmentation.

To clarify, 'octupling the input data' refers to the process of increasing the size of our training dataset by a factor of eight through data augmentation. This technique involves applying a series of basic image manipulations, such as flipping horizontally, flipping vertically, rotating, shearing, and scaling, to each image in our original dataset. These manipulations generate multiple variations of each image, thereby expanding our dataset and enhancing the diversity of training examples. This strategy is critical for improving the neural network's generalization capabilities and reducing the risk of overfitting, especially given our deliberate use of a smaller initial dataset.

To ensure comprehensibility, we propose to slightly rephrase the sentence in the paper to:” 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 – i.e., generating eight variations of each original image – in order to increase its initial size rendering the neural network more robust, while at the same time compensating for deliberately using a small input dataset.”

This revision aims to succinctly explain the concept of data augmentation and its purpose in our study, making it more accessible to readers who may be less familiar with these machine learning practices.

Line 181 and Figure 3: it would help if in the figure you can indicate which grains are celestite.
also why is the red of bassanite in fig 3e more like an orange whereas in d and c it is more like a faint red like in the histogram? the color in the legend does not correspond to the shaded area in the histogram, it is a different shade at least in this figure. please be consistent

The celestite is already labelled in Figure 3a and we prefer not to add extra text risking cramming the figure. We have now fixed the colours scheme in Fig. 3e.

Line 191. how do the various filter affect the output of the random forest? are they applied all toegther or separately and then the one that receive the most votes is chosen?

Thank you for your question. 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.

We have now provided these details in section “Appendix B: Random Forest Segmentation”

Figure 3 caption: “cross-sectional” instead of “Horizontal”

Modified as suggested

Line 215: indicating that is the best result?

Yes, values closer to 1 indicate a better scoring and that the outputs produced by the deep learning model is closer to the labelled input the user has provided.

Line 224 – 226: I would rephrase this sentence with "the model trained using a ground truth from a random forest classifier was the only model producing a DICE score of 0.98 for all phases." to make it clearer.

This is not entirely true. When using the RF model, the DICE score is not 0.98 for all phases (as shown in Table 1). 0.98 is the max value reached by the DICE score, specifically for “Gypsum DICE”. The main advantage of RF over Histogram segmentation is that RF is the only model able to produce a score (and therefore correctly identify) for the Celestite phase.

Line 230: why should they not be consecutive?!

We could have chosen to test the model over a set of not consecutive slices cropped from the µCT volume.

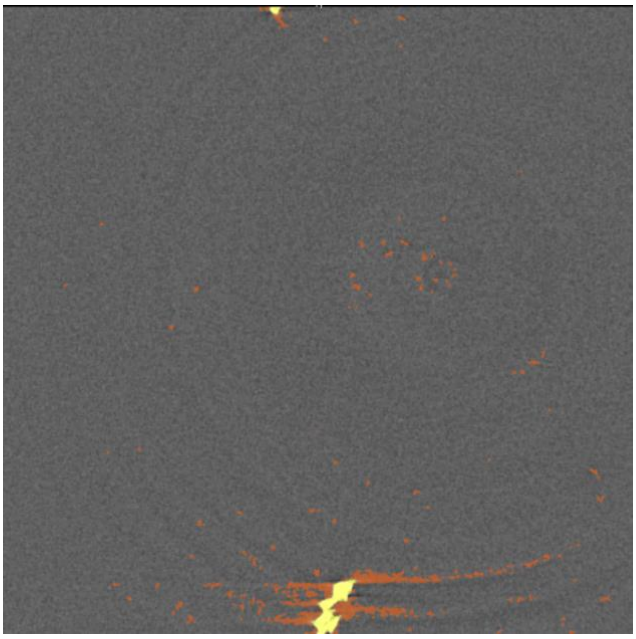
Line 234: can the model be applied as batch-process using some IT cluster?

Yes, the model can indeed be run outside the Dragonfly platform for batch processing on clusters. However, this would involve implementing a series of scripts or code that perform the same steps currently executed within Dragonfly. These steps include data pre-processing, applying the model to the data, and handling the output. While this requires additional effort to set up the necessary computational environment and codebase, it is certainly feasible and allows the model to be utilised in a more flexible and scalable manner.

Figure 5: the results are rather astonishing however I can see for this celestite grain that celestite has been classified where the original image is grey (middle of the grain). Is it possible to see the same grain after the post-processing where errors should have been cleaned up and mislabel pixels fixed? it would be good to see all steps for the same image so including the post-processing. or do you reckon those grey pixels are actually caused by something else? (and what)

The images shown in figure 5 are already post-processed and post-processing does not involve the celestite. As detailed in the paper the main targets of post-processing are mislabelled pixels of Gypsum interpreted as

Bassanite during the early stages of dehydration. Mislabelling occurs because at early dehydration stages the S-μCT images are inherently noisy due to the homogeneity of the gypsum phase. I have attached a figure here that exemplifies the issues.



Line 251: I would like to see an example of that. It would be even better if it can be on the previous figure you showed.

See picture above.

Line 257: why is it minor? not clear to me from this sentence

The mislabelling problems are encountered in μCT volumes in the early stages of the dehydration and vanished as soon as the Bassanite phase begins to grow, reducing the homogeneity in the μCT.

Line 298 and Figure 7: are the data points in Fig 7 relative to the same time scan? and if so, how do we know how much reaction % there should be? Perhaps I am reading the figure in the wrong way.. in a), the second data point from the left measured an approx. 27% reaction %, whereas in b) is less than 20% which is quite a big difference. can you clarify why a) is more accurate based on this graph?

I find this figure hard to read, even with the caption.

Figure 7 presents a quantitative analysis of phase volume changes during the gypsum dehydration experiment, specifically focusing on the same time series data (volume VA19) segmented using two distinct methods. The first method (represented by dots in 'a') utilizes the workflow developed in this study, combining a random forest classifier with a deep learning model. The second method (represented by diamonds in 'b') employs a conventional thresholding segmentation strategy.

The discrepancy in reaction percentages between these two approaches, particularly noticeable in the celestite phase, highlights the enhanced accuracy of our developed workflow. Unlike conventional thresholding segmentation, our method can more precisely detect and quantify all phases, including those that are typically challenging to segment like celestite. This is demonstrated by the more consistent and accurate volume measurements obtained with our workflow compared to the underestimation of reaction extent by the conventional method.

Therefore, the differences in reaction percentages as observed in Figure 7 underscore the superior segmentation capabilities of our Random Forest + Deep Learning approach, particularly in accurately capturing phases that are not discernible using standard histogram thresholding methods.

Line 310 – 311: this last sentence is unclear or missing something.

Thanks for picking this up... there was a typo in the manuscript. This is now: “To ascertain the accuracy of the chosen deep learning model we compared the theoretical and measured molar evolution of gypsum to bassanite during dehydration”

Line 333: why is hyperparameters written like that, and earlier on called (hyper)-parameters? we still do not have a definition for what these (hyper) parameters are and why they are just not called parameters.

Thanks for flagging this typo. Now fixed.

Line 340: repetition ot the sentence... : “These external standards aid in the assurance of measurement accuracy [Writers et al., 2021].”

Thanks for pointing this out. This has now been fixed.

Line 349: what if there is no prior knowledge of the chemical reaction involved? what would be an alternative to check segmentation accuracy?

Thank you for your insightful question. Indeed, the a-priori knowledge of chemical reactions plays a crucial role in establishing a framework for assessing the accuracy of data extracted from µCT images, as mentioned in Section 4.1 of our paper.

As highlighted in the Discussion section of our paper, one common strategy in fields like medical imaging and material science is the use of external calibration techniques. These techniques involve employing phantoms—

objects with known dimensions and/or compositions—as benchmarks. Phantoms serve as external standards that can be imaged and segmented to evaluate the accuracy of segmentation methods. By comparing the segmented volumes or dimensions of these phantoms with their known actual values, we can assess the accuracy and reliability of segmentation techniques used in µCT imaging.

Additional alternative strategies that can be consider for benchmarking segmentation outputs are: (a) Comparative Studies: Conducting comparative studies with existing, well-validated segmentation algorithms can also serve as a validation technique. If the new segmentation results are in good agreement with those obtained from established methods, it adds credibility to the new method. (b) Synthetic or Simulated Data: Using synthetic or simulated datasets where the true segmentation is known can be a powerful tool for validation. By applying the segmentation algorithm to these controlled datasets, its accuracy and reliability can be rigorously tested.

Line 377: this sentence is not grammatically correct, please revise.

Thanks for the suggestion. Now: "Future iterations of our method will aim to expand its capabilities and applications."

Line 380: what is transfer learning and reinforcement learning? not clear from the follow up sentences.

Thank you for your question. In the text we have provided few references when mentioning these methods and for the scope of our paper we think that pointing the interested readers to these is sufficient. However, since also Reviewer 3 has raised the issue that this paragraph was unclear, we have now integrated the text as follows: "Future iterations of our method will aim 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."

Line 385: it would be good to explain how supervised and unsupervised methods differ since you talked about both concept in this paper.

As detailed in a previous response we have now integrated in the text the difference between the two approaches.

Line 387: please connect these two sentences, or if using two separate sentences do not use "while" at the beginning of the sentence

We have now revised the paragraphs as follows: "Unsupervised learning can dramatically reduce the time and effort required for data annotation, thereby accelerating analysis, and enabling the exploration of larger datasets (Mahdaviara et al., 2023). Additionally, leveraging data from before and after a scan in a time series can provide extra information, further enhancing our ability to segment complex datasets more effectively."

<u>Response to Reviewer #2</u> - Richard Ketcham

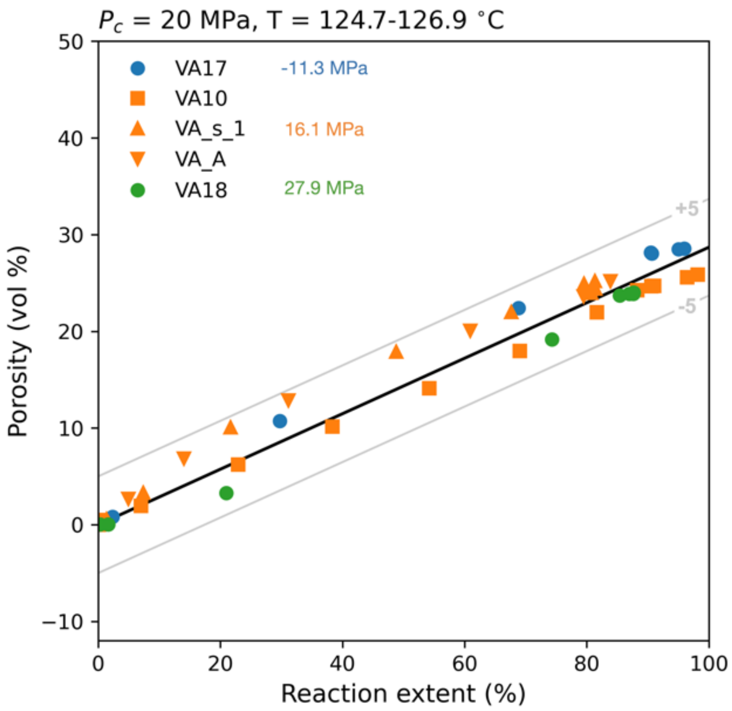
This is a very nice study, using knowledge of the physical system as a means of benchmarking machine learning approaches to segmentation. It will certainly be of interest to the community, and is close to being ready to publish as-is. There are only a few items that might need a little more work.

We thank Prof. Richard Ketcham for the comments and suggestions on our manuscript. We are pleased to read that he found this work valuable contribution. We hope our response to the comments addresses the issues raised.

Overall this method seems to work quite well on this system. However, it relies on manual segmentation to create the training data. This works fine for the more massive phases, but for the ones that have one or more small dimensions relative to the data (celestite, and more importantly water-filled porosity) many of the voxels suffer partial-volume or blurring effects. The segmentation of the porosity in Fig. 3e is very chunky in comparison to how the porosity really is (thin grain boundary layers). Basically, everything with a hint of darkness is being called porosity, but many of those voxels represent mixtures between water and solid material. This makes the method a bit less repeatable, and is likely why the water always comes out a bit high in Fig. 7. This issue is recognized by the authors in lines 341-345, but not given context on how it is affecting the analysis presented.

We acknowledge the limitations of our segmentation method. The manuscript demonstrates the precision of segmentation by comparing it with theoretical molar volumes (please refer to Figure 7). It is shown that the deviation from theoretical values for porosity—representing water in the pore space—falls within a 95% accuracy threshold (as evidenced by the proximity of the data points to the theoretical lines in Figure 7a).

To maintain focus and clarity in the paper, we have presented the accuracy analysis for only one time-series volume. However, in a recently accepted publication in Geology (Gilgannon et al., 2023), we have demonstrated the consistent accuracy across all volume data. The accompanying image below illustrates the extracted porosity from various experiments (represented by dots, triangles, diamonds, and squares) alongside the theoretical curve for the dehydration reaction



Also, the calibration of CT values does not necessarily need to rely on external calibration – you can probably just measure the grayscales off your images (and use the volume balance in your system to check that your results are about right). Since this would essentially be a post-processing step reinterpreting some segmented voxels, it seems a complementary add-on.

We are grateful to Prof. Ketcham for his insightful comment on calibrating µCT values. The recommendation to derive grayscale values directly from our µCT images and utilise the volume balance of our system for validation is well-received. We will explore how to incorporate this method into our existing procedures, mindful of our dataset's unique characteristics and constraints. While implementing this technique may pose challenges, particularly in matching grayscale values to specific phases or compositions, it represents a promising direction for advancing and fine-tuning our segmentation approach.

Detailed comments:

[line 72] Replace "results" with "provides"

Modified as suggested.

[lines 163-165] This sentence is garbled ("histogram thresholding… or… histogram thresholding"?). Rewrite.

This issue was also raised by Reviewer 1. We have, therefore, revised the sentence as follows: "To establish the ground truth set for initial image classification, we initially considered methods with differing informational depths. One such method is histogram thresholding, which relies on basic greyscale values and typically results in low-information-level outcomes. However, this approach alone proved inadequate for our purpose, as it often led to gradients with diffuse phase boundaries, underscoring the need for more sophisticated classification methods."

[Table 1] It seems odd that the average DICE for Model E is higher than that for Model RF even though the former scores worse and much worse on two phases. How was the average calculated?

The average is derived from only those classes that obtain a DICE score. To provide a clearer picture, as average DICE values may not fully represent individual class performance, we present the DICE scores for each segmented class in the table.

[line 314] What does "(imaged as porosity [in] the µCT data)" mean? Was the water missing/drained, or is the water just being called porosity because it's pore-filling?

A portion of the water released during the reaction was retained in the porosity formed from the volumetric reduction when gypsum dehydrated to bassanite, while the surplus was expelled. Given that the fluid pressure (Pf) applied surpassed the vapor pressure of water at the reaction temperature, it is likely that all imaged porosity was water-saturated.

[line 331] Another issue with ML-segmentation is that the training may only work well for very similar conditions of material, geometry, imaging parameters, etc.

We fully agree. This is the reason why we implemented a Deep Leaning (DL) model on top of the ML Random Forest. The DL is able to generalise a segmentation model better than ML, particularly when DL is integrated with Data Augmentation.

[line 344-345] Another approach might be re-interpretation of voxels using grayscale, so as to assign affected voxels partial values.

We have now included this into the Discussion session as follows: "Solutions have been proposed and often require complementary techniques (such as using tactile, optical sensors) to calibrate measurements derived from CT data (Torralba, 2018). Reinterpreting segmented voxels using grayscale values can offer a complementary method for calibration. This approach assigns partial values to affected voxels, potentially enhancing accuracy in cases of overlapping mineral phases or partial volume effects."

[line 377] “In future iterations of our method aims to expand…”: change to “we aim” or “Future iterations of our method will aim”

Thanks for the suggestion. Now: “Future iterations of our method will aim to expand its capabilities and applications”

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.

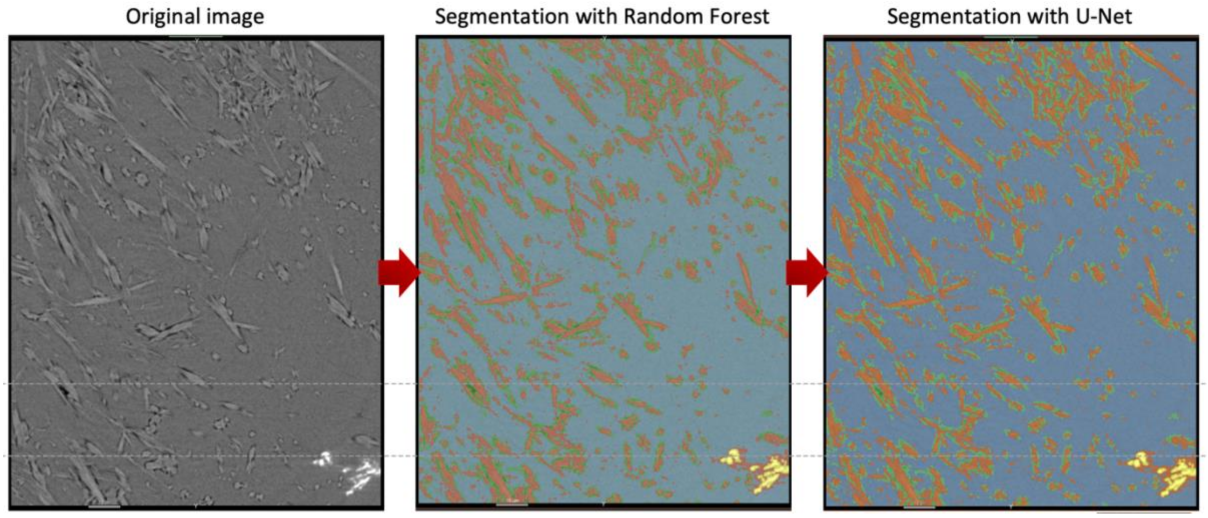
- How was the validation set chosen? Equally across all images from each time step? I would be interested to see how the model performs if the validation set contained all images of a given time step (therefore not present in the training set). As it is argued that the DL method is more robust and consistent,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