ASME Journal of Nondestructive Evaluation, Diagnostics and Prognostics of Engineering Systems
Online journal at:
https://asmedigitalcollection.asme.org/nondestructive

Evaluation of Flaw Detection Algorithm Using Simulated X-Ray Computed Tomography of Ground Truth Data

Felix H. Kim[1]
National Institute of Standards and Technology,
Gaithersburg, MD 20899
e-mail: felix.kim@nist.gov

Adam L. Pintar
National Institute of Standards and Technology,
Gaithersburg, MD 20899
e-mail: adam.pintar@nist.gov

John Henry J. Scott
National Institute of Standards and Technology,
Gaithersburg, MD 20899
e-mail: johnhenry.scott@nist.gov

Edward J. Garboczi
National Institute of Standards and Technology,
Gaithersburg, MD 20899
e-mail: edward.garboczi@nist.gov

A framework to generate simulated X-ray computed tomography (XCT) data of ground truth (denoted here as "GT") flaws was developed for the evaluation of flaw detection algorithms using image comparison metrics. The flaws mimic some of those found in additively manufactured parts. The simulated flaw structure gave a GT data set with which to quantitatively evaluate, by calculating exact errors, the results of flaw detection algorithms applied to simulated XCT images. The simulated data avoided time-consuming manual voxel labeling steps needed for many physical data sets to generate GT images. The voxelated pore meshes that exactly match GT images were used in this study as opposed to using continuum pore meshes. The voxelated pore mesh approach avoids approximation error that occurs when converting continuum pore meshes to voxelated GT images. Spherical pores of varying sizes were randomly distributed near the surface and interior of a cylindrical part. XCT simulation was carried out on the structure at three different signal-to-noise levels by changing the number of frames integrated for each projection. Two different local thresholding algorithms (a commercial code and the Bernsen method) and a global thresholding algorithm (Otsu) were used to segment images using varying sets of algorithm parameters. The segmentation results were evaluated with various image evaluation metrics, which showed different behaviors for the three algorithms regarding "closeness" to the GT data. An approach to optimize the thresholding parameters was demonstrated for the commercial flaw detection algorithm based on semantic evaluation metrics. A framework to evaluate pore sizing error and binary probability of detection was further demonstrated to compare the optimization results. [DOI: 10.1115/1.4063170]

Keywords: X-ray computed tomography, defect, flaw, additive manufacturing, image segmentation, evaluation metrics, ground truth data, nondestructive evaluation, imaging, testing methodologies

1 Introduction

X-ray computed tomography (XCT) is a powerful tool for nondestructive detection evaluation of flaws in traditionally processed materials, as well as in additive manufacturing (AM) materials [1]. AM parts can include volumetric flaws such as gas pores, lack-of-fusion pores, keyholing pores, near-surface pores, cracks, and delamination [2–6]. XCT scans can be used to characterize the pores [7] or to detect and screen out defective parts based on a probability of detection (POD) curve [8]. For quantitative measurements/characterization of pores and automated/assisted flaw detection, a flaw detection algorithm is generally implemented through image thresholding [9] or machine learning [10,11]. Regardless of the algorithm, the accuracy of the algorithm and associated parameters used are often difficult to evaluate due to a lack of ground truth (GT) information. The GT information may be obtained through destructive measurements, which are often time consuming and tedious, and the part is destroyed after the process and is not reusable. It would be useful to have a reusable physical phantom or easily configurable synthetic phantom (we use the term *phantom* instead of artifact in this article so as not to confuse the reader with well-known XCT imaging *artifacts* such as beam hardening and scattering).

There is, however, a lack of physical or synthetic phantoms to evaluate flaw detection algorithms. Existing physical phantoms were mostly developed to measure spatial resolution [12,13]. Synthetic phantoms were developed for the evaluation of reconstruction algorithm image quality for medical imaging applications (e.g., the Shepp–Logan phantom [14]). Simulated phantoms for scientific applications with synchrotron imaging were also developed to assess reconstruction algorithms [15]. Similarly, a software package was developed to generate application-specific phantoms such as those for dynamic imaging [16]. Pore detection and measurement capabilities have not been assessed using these phantoms. This work addresses

[1]Corresponding author.
Manuscript received December 10, 2022; final manuscript received August 9, 2023; published online October 4, 2023. Assoc. Editor: Kara Peters.

these deficiencies by focusing on the development of synthetic phantoms specifically for the evaluation of flaw detection algorithms.

State-of-the-art XCT simulation codes can generate realistic simulated data [17], and this capability allows the development of realistic synthetic phantoms. Synthetic phantoms have two clear benefits: (1) the flexibility of developing custom part geometries and pore distributions and (2) a clear knowledge of the GT information regarding the original pore sizes, shapes, and locations. Both benefits result from the fact that the synthetic phantom is created voxel-by-voxel, inputting known information into every voxel. By using these synthetic phantoms, we know *exactly* what the binarized image should be after segmentation.

Porosity is a measure of the total amount of void space in a material, which can be measured using different techniques [18]. In fracture mechanics, however, the individual pore characteristics (size, shape, and location) are of interest [19]. Due to the challenges of measuring individual pore characteristics using traditional measurement techniques, the total porosity is often used to estimate the mechanical properties and structural integrity of the sample. XCT, on the other hand, allows measurements of individual pore characteristics from the flaw masks (i.e., segmented images) identified by a detection algorithm, and evaluating the accuracy of these measurements is the focus of this article.

In this article, we demonstrate a workflow to generate realistic XCT simulation data with spherical pores placed at locations generated from appropriate probability distributions and evaluate various segmentation tasks. The main purpose of developing the synthetic phantom and simulated images is to provide benchmark data sets for comparison between different segmentation algorithms. We demonstrate the concept on a standard cylindrical part geometry, but our approach can be readily applied to different part geometries. While similar XCT simulation-based approaches were independently implemented to train and evaluate machine learning-based flaw detection algorithms [20] and to demonstrate a digital twin framework [21], we implemented a unique approach to create unambiguous GT labeled data based on voxelated pores. Although it is possible to generate labeled GT images based on the amount of overlap of surface meshes with image grids (i.e., discretization), our approach ensures that the GT image was exactly what was used to generate the simulated XCT images. We further investigate a comprehensive list of metrics to evaluate semantic segmentation and instance segmentation of individual pores (pore-level size errors and binary POD curves) to compare performance of different segmentation algorithms (and different algorithm parameters) and discuss the differences between semantic segmentation evaluation metrics and pore-level evaluation metrics. We also demonstrate an approach to optimize local thresholding algorithm parameters using the generated data sets for several evaluation metrics and discuss transferability of information (e.g., optimized algorithm parameters and measurement uncertainty) from simulated data sets to similar experimental data for calibration of flaw sizes. The methods demonstrated in this article can be applied to evaluate accuracy and reliability of different detection algorithms, which is useful for the NDE 4.0 framework [22].

2 Data Generation

The evaluation data were generated through XCT simulation using a commercial simulation code (aRTist[2]) developed by the German Federal Institute for Materials Research and Testing (BAM) [23]. Our synthetic phantom, similar to that used for a detailed pore characterization study [24], was a cylindrical specimen (approximately 3 mm diameter × 3 mm height). The data generated in this article are publicly available [25].

[2]Certain commercial equipment, instruments, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

2.1 Ground Truth Image Development. A cylindrical matrix (827 voxel diameter × 827 voxel height) was generated in the center of a 1001^3 voxel structure with 3.63 μm/voxel equivalent to XCT simulation voxel resolution. A total of 420 nonoverlapping pores were randomly distributed in the part representing both near-surface and volumetric spherical gas pores. Pore diameters ranged from 1 voxel to 41 voxels (diameters of 3.63 μm to 148.8 μm) in steps of two voxels (7.26 μm). Smaller pores (<7 voxel diameter) were closer to a cubic shape, but for voxel diameters greater than about 7, the pore volume agreed with the true continuum sphere volume within 2%. Ten pores of each size were randomly distributed in the near-surface region within 200 μm of the sides of the cylinder, and ten pores of each size were randomly distributed in the rest of the article. To avoid placing pores too close to the surface, the minimum distance between any part of a pore and the cylindrical part surface was limited to 10 voxels (36.3 μm). It was out of scope to study the effect of pores being any closer to the surface, whose detection results would have been influenced by software's surface determination capability. Horizontal and vertical cross section examples are shown in Fig. 1. This synthetic phantom served as the GT image for the XCT simulation and the evaluation of the segmentation algorithms. Different pore sizes were labeled by different shades of gray in Fig. 1, for ease of visualization and identification, but all pores were binarized for the analysis described in this article. While larger numbers of pores and additional variations of pore characteristics (size, shape, orientation, and distribution) can be implemented, we do not do so here to provide simple, clear GT data for evaluation. We plan to incorporate additional complexities in the future, as discussed later.

2.2 Pore Mesh Generation. The XCT simulation was based on surface meshes and associated material properties. We generated voxelated surface meshes that followed the boundaries of the pore voxel structure developed in the previous step. The voxelated pore surface meshes were generated by using voxelfuse [26], and duplicate faces were removed using meshlab [27]. Although pores of continuum shape can be generated (Fig. 2(*a*)), we opted to directly generate voxelated pore meshes (Fig. 2(*b*)). This avoided discretization of a continuum surface mesh to a voxel grid in order to generate GT images (Fig. 2(*c*)). This approximate conversion is a source of uncertainty in the voxel-to-voxel evaluation process. In our case, the voxelated mesh used in the XCT simulation perfectly matched the voxelated GT image (Fig. 2(*d*)), so the segmentation from the XCT reconstruction can be directly compared to the GT image. Examples of voxelated surface meshes are shown in Fig. 2(*e*).

2.3 X-Ray Computed Tomography Simulation. In the XCT simulation software, a cylinder with a diameter of approximately 3 mm and a height of 3 mm was generated with the material properties of nickel super alloy 625 (IN625). The voxelated surface meshes of the pores were placed in the cylinder according to their positions determined in the previous step. The parameters shown in Table 1 were used for the simulations. Camera binning and optical magnification were considered when selecting the effective detector pixel pitch. The detector size was 1001 × 1001 pixels, which generated 1001 × 1001 × 1001 voxel volumes after reconstruction identical to the GT image size. The detector model was used to account for the image intensity and the signal-to-noise ratio (SNR) of a radiograph acquired at a specified distance from the source and for a specified exposure time and number of frames integrated. This allowed estimation of image intensity and the SNR at different imaging conditions from reference measurements. In this example, we used 1, 4, and 9 frame averages, which effectively gave normalized SNR values of 1×, 2×, and 3×, respectively. The parameters used in the detector model are provided in the Supplemental Materials. The effects of X-ray scattering (Compton and Rayleigh) were also considered by implementing the Monte Carlo-based McRay simulation [28] at every ten projections (2.25 deg) using 10^7 photons. The effect of X-ray scattering was found to be minimal under the current test conditions. A spot size of 5 μm was used.

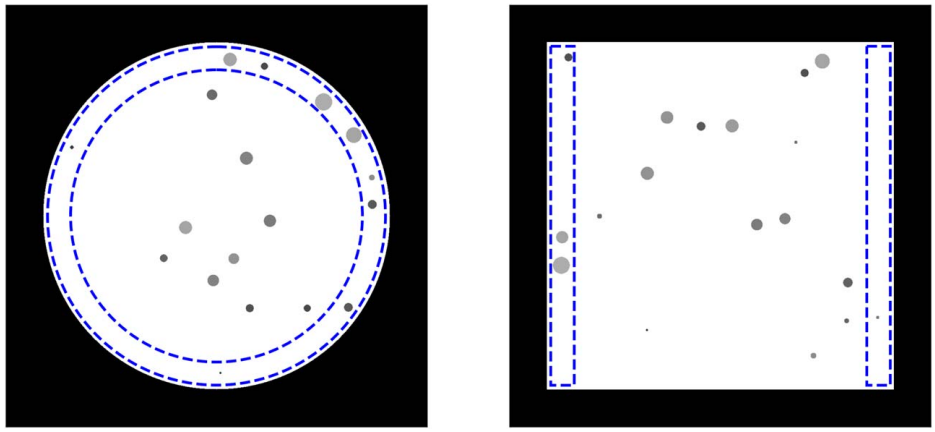

Fig. 1 A horizontal cross section (left) and a vertical cross section (right) of the phantom. The near-surface region is shown as the dashed area

2.4 X-Ray Computed Tomography Reconstruction. XCT reconstruction was carried out using the open-source software ASTRA [29,30], which uses the Feldkamp, Davis, Kress (FDK) algorithm [31]. The effects of iterative algorithms will be further investigated in future studies. The original reconstructed attenuation values are in units of mm^{-1}, and the values were rescaled to 16-bit [0–65,535] grayscale values (unsigned integers) in a consistent manner for all data sets. The XCT reconstructed image intensity (floating point number) is related to the attenuation coefficient in a unit of length set by the user (mm^{-1} in floating point precision). We linearly rescaled the reconstructions to 16-bit unsigned integers by setting the peak of the air grayscale values to be approximately 13,000. The peak grayscale value of the material was set at 50,000 (slope $= 5.405 \times 10^{-5}$ and offset $= -0.703$). The slope and offset can be used to convert the 16-bit unsigned integers back to floating point values, up to round-off error with digital noise, if needed (floating point value = 16-bit grayscale value × slope + offset). Examples of cross-sectional reconstructed images with different SNRs are shown in Figs. 3(*a*)–3(*c*), and small regions of interests (ROIs) are compared in Figs. 3(*d*)–3(*f*). The SNR values are local values achieved within the ROIs shown, and they can vary depending on the ROI locations and sizes. Grayscale histograms in Fig. 3(*g*) show the effect of different SNRs. The histogram peaks are well separated from each other, and the peaks get sharper as the SNR improves. The image intensity values are shown for different sizes of pores and solid material in Fig. 4. The voxelated pore masks were conveniently used to find image intensities of pores and solid material, and the local contrast (difference in the image intensity of solid and pore) increased as the pore size increased. The average image intensities (points) and their ±1 standard deviations (error bars) are plotted for different pore sizes, and the solid material image intensity (constant horizontal line) ± 1 standard deviation (constant error bars) are also plotted. In this case, the solid mask was eroded by using a spherical structural element (61-voxel diameter) to measure the intensity and noise values sufficiently away from solid/pore interfaces. The average values of pore intensity are similar for different SNR data, while a slight reduction in standard deviation is found for the increase in SNRs. The variability of image intensity in solid material voxels is affected by the presence of pores; therefore, SNR values were evaluated in regions sufficiently far away from these pores.

3 Image Segmentation

Image segmentation was the basis for feature detection in our XCT images by creating a binary mask through classification of the voxels as black (pore) or white (solid material). Segmentation usually involves the implementation of an image thresholding algorithm combined with some preprocessing (e.g., de-noising) and postprocessing steps (e.g., morphological operations). We implemented a commercial local thresholding algorithm (EasyPore algorithm in VGStudioMax 3.4), the Bernsen local threshold algorithm [32] implemented in 3D, and the Otsu global thresholding algorithm [33]. Both local thresholding algorithms have two main

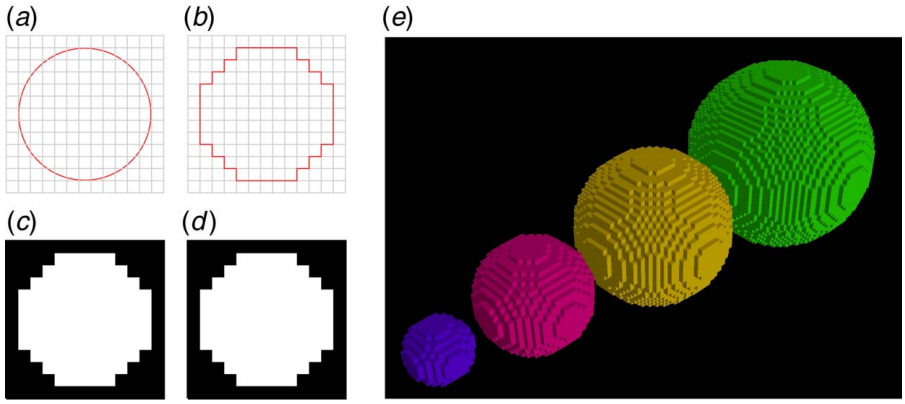

Fig. 2 Surface meshes that can be used for XCT simulation with (*a*) continuum shape and (*b*) voxelated structure. (c) and (*d*) Corresponding binary GT images of (*a*) and (*b*). (e) Voxelated surface meshes (≈11, 21, 31, and 41 voxel diameters).

Table 1 XCT simulation parameters

Parameters	Values
Voltage (kV)	160
Current (μA)	61
Target	W
Window (material; thickness (mm))	Be; 1
Filter (material; thickness (mm))	Mo; 5
Exposure time (s)	3
Frames/projection	1, 4, 9
Number of projections	1600
Source-to-detector distance (mm)	26
Source-to-object distance (mm)	14
Geometric magnification	1.86
Camera pixel pitch (μm)	13.5
Binning	2
Optical magnification	4
Effective detector pixel pitch (μm)	6.742
Effective voxel pitch (μm)	3.630

tunable parameters: the local contrast threshold (LCT) and local window (LW) size, which are specified by the user. A local thresholding algorithm generally determines the threshold based on information within each local window surrounding the voxel to be classified [9], which tends to outperform global thresholding algorithms when there is nonuniformity in image intensity [34]. The outside surface of the metal cylinder (part surface) was determined using the commercial software at the ISO45 (45% value between the histogram peaks of material and air, see Fig. 3(*g*)) value initially and further refined iteratively. For the Bernsen algorithm, a minimal follow-up image processing step was implemented to mask the boundary of the region to be studied. Inaccuracy in the surface determination can inadvertently penalize the evaluation of flaw detection algorithms by omitting some pores near surfaces within this bounding volume. To avoid this, we ensured that all pores were present within the bounding volumes used. For the Bernsen algorithm, the thresholding was initially carried out on the entire image, which included background areas. In order to only account for the features detected in the part of interest, a cylinder that was eight voxels smaller than the original cylinder was used as a mask. The radius of the largest local window size considered in this study was also eight voxels, which essentially limited how close to the edge one can evaluate the results using the local thresholding algorithm. The Otsu algorithm does not have any user-specified parameters, but it automatically finds $n-1$ global thresholds between n histogram peaks ($n=2$ in Fig. 3).

Examples of the voxelated pore mesh used for the simulation, GT image, reconstructed image, and segmented image are shown in Figs. 5(*a*)–5(*d*), respectively. The voxelated pore surface meshes used for the XCT simulation are overlaid on the binary GT image, reconstructed image, and thresholded image. The surface mesh is the same size as the GT image pore boundary so that the GT image can be directly compared with the segmented image. In this example, the segmented pore was slightly larger than the GT pore based on the LCT chosen for demonstration purposes.

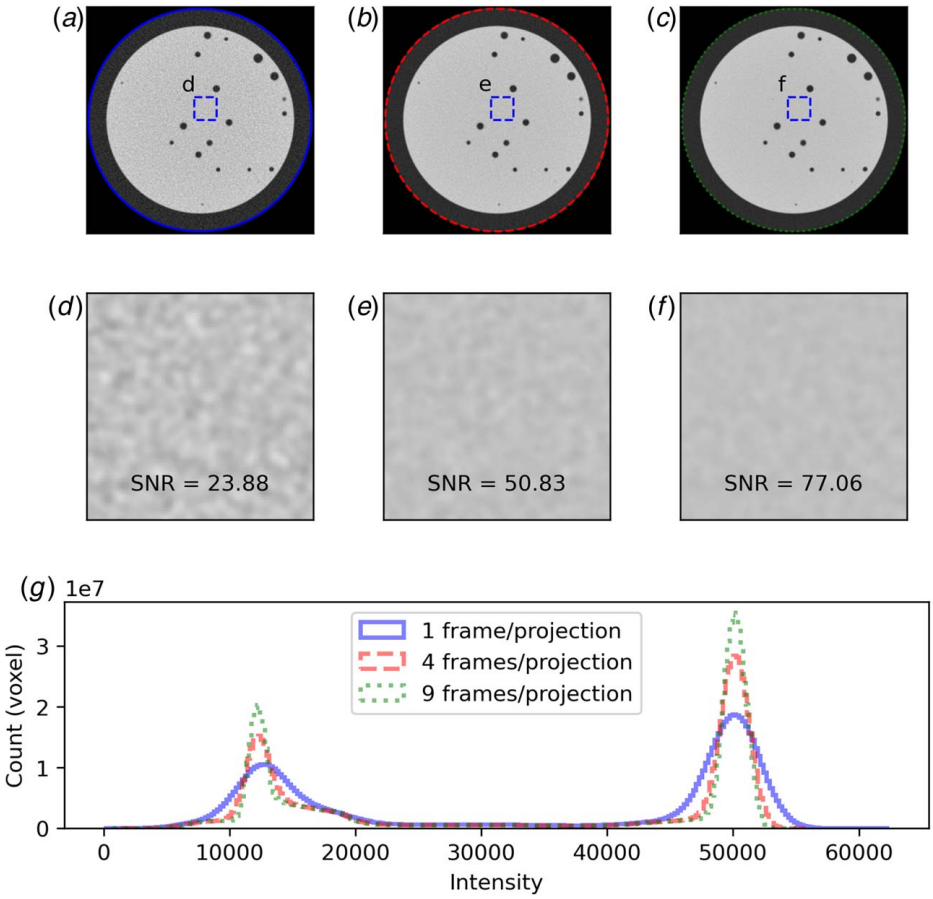

Fig. 3 Example reconstructed image of simulated using (*a*) 1 frame/projection (fpp), (*b*) 4 fpp, and (*c*) 9 fpp. The small squares indicated by a dashed line are shown in (*d*) 1 fpp (from (*a*)), (*e*) 4 fpp (from (*b*)), and (*f*) 9 fpp (from (*c*)). (*g*) Comparison of the grayscale histograms for each image (*a*, *b*, and *c*).

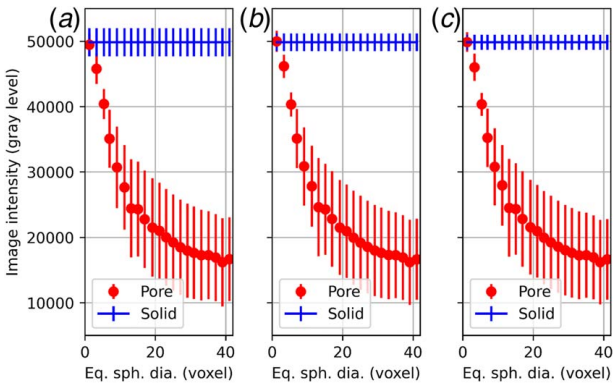

Fig. 4 Image contrast of pores and solid material for (*a*) 1 fpp, (*b*) 4 fpp, and (*c*) 9 fpp. Error bars are ±1 standard deviation

4 Evaluation Metrics

The segmentation results were evaluated globally (semantic segmentation evaluation) for an entire 3D image, using a wide range of metrics in order to maximize the usefulness of this evaluation, and locally for individual pores. The metrics and methods used to evaluate the detection results are described in this section.

4.1 Semantic Segmentation Evaluation. The semantic segmentation evaluation was carried out by comparing the segmented image to the GT image using metrics or measures of similarity or distance between the images. The metrics in Table 2 were used to evaluate the segmentation results.

The equation for intersection over union (IOU) is shown in Eq. (1) as it will be used throughout the article. The symbols A and B represent two images being compared, and many metrics are based on components of a confusion matrix, which include the true positive counts (TP, pixels/voxels that are correctly segmented as flaws), false positive counts (FP, pixels/voxels that are incorrectly segmented as flaws), true negative counts (TN, pixels/voxels that are correctly segmented as solid material or background), and false negative counts (FN, pixels/voxels that are incorrectly segmented as solid material or background). Basic equations for other metrics are provided in the Supplemental Materials of this article and summarized in Ref. [53]. The metrics can be grouped into six categories: (1) overlap based (DICE, IOU, accuracy (ACR), true positive rate (TPR), true negative rate (TNR), false positive rate (FPR), false negative rate (FNR), precision (PRC),

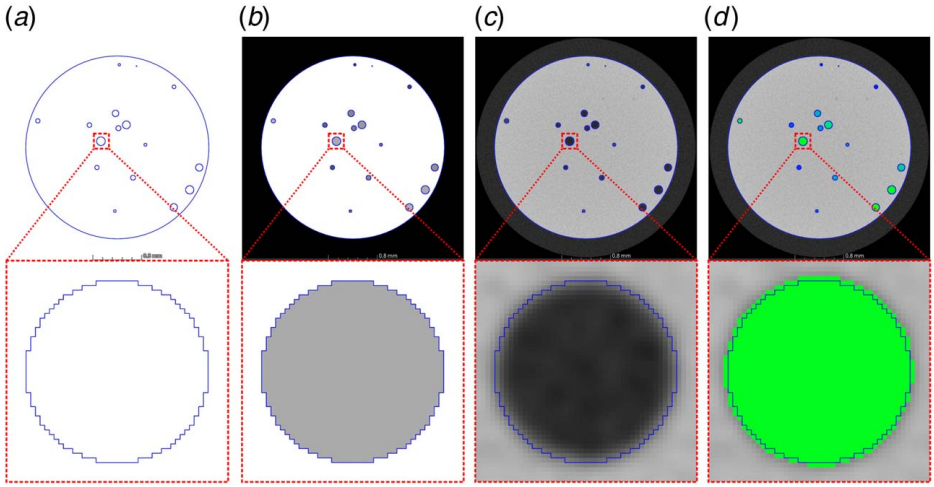

Fig. 5 (*a*) Pore surface mesh used for simulation, (*b*) GT image, (c) reconstructed image, and (*d*) segmented image. The saw-tooth line is the surface of the GT particle (in this 2D slice)

Table 2 Metrics and symbols used in this article

Metric	Symbol	Other names	Category	References
Dice	DICE	=F1-Measure	Overlap based	[35]
Intersection over union	IOU	Jaccard index	Overlap based	[36]
Accuracy	ACR		Overlap based	
True positive rate	TPR	Sensitivity, Recall	Overlap based	
True negative rate	TNR	Specificity	Overlap based	
False positive rate	FPR	Fallout, = 1 – TNR	Overlap based	
False negative rate	FNR	=1-TPR	Overlap based	
F-measure	FMS	F1-measure, = Dice	Overlap based	[37]
Precision	PRC		Overlap based	
Global consistency error	GCE		Overlap based	[38]
Volumetric similarity	VS		Volume based	[39–42]
Rand index	RI		Pair-counting based	[43]
Adjusted Rand index	ARI		Pair-counting based	[44]
Mutual information	MI		Information theoretic based	[45,46]
Variation of information	VOI		Information theoretic based	[47]
Interclass correlation	ICC		Probabilistic based	[48]
Probabilistic distance	PBD		Probabilistic based	[49]
Cohens kappa	KAP		Probabilistic based	[50]
Area under ROC curve	AUC		Probabilistic based	[51]
Mahalanobis distance	MHD		Distance based	[52]

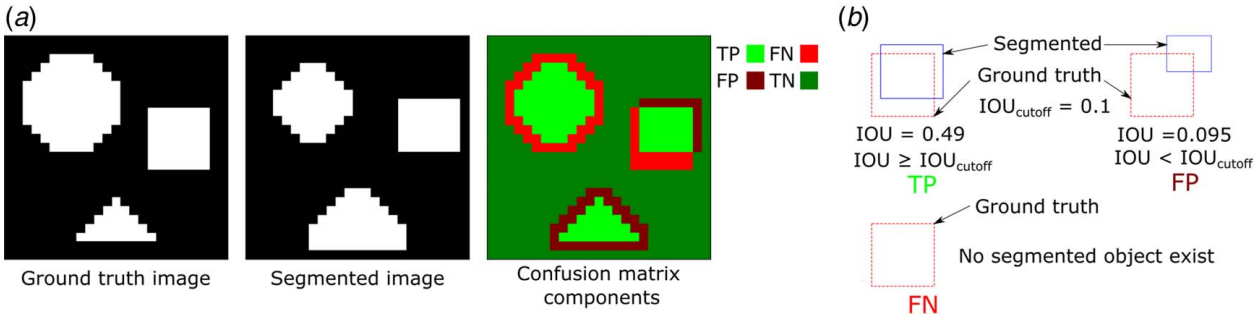

Fig. 6 (*a*) Illustration of semantic segmentation evaluation and confusion matrix components for semantic segmentation and (*b*) illustration of instance segmentation and object detection

F-measure (FMS), global consistency error (GCE)), (2) volume based (volumetric similarity (VS)), (3) pair-counting based (Rand index (RI), adjusted Rand index (ARI)), (4) information theoretic based (mutual information (MI), variation of information (VOI)), (5) probabilistic based (interclass correlation (ICC), probabilistic distance (PBD), Cohens kappa (KAP), area under receiver operating characteristic (ROC) curve (AUC)), and (6) spatial distance based (Mahalanobis distance (MHD)). The AUC for binary image evaluation follows the definition found in Ref. [51] by measuring the areas under trapezoids for a single-point ROC curve connecting (0,0) and (1,1). The MHD metric follows the definition of [52]. For the type of evaluation carried out in this article, DICE is identical to FMS. The IOU metric is also called the Jaccard index, and it is related to DICE (=2*IOU/(1 + IOU)). Therefore, DICE and IOU provide similar information. The TPR is also called sensitivity, TNR is also called specificity, and FPR is also called fallout. The FPR is related to TNR (FPR = 1–TNR), and FNR is related to TPR (FNR = 1–TPR). Therefore, either FPR or TNR and FNR or TPR is recommended to be used for evaluation. A simple graphical illustration is shown in Fig. 6(*a*) for semantic segmentation evaluation and confusion matrix components. The EvaluateSegmentation tool [53] was used to compute all semantic segmentation evaluation metrics. While we investigated all metrics listed in Table 2, we found that only a few were relevant to this article. It is worth noting that different metrics are sensitive to different properties for evaluation, and the choice of metric is considered to be application specific [53]. We intend to apply these metrics for the evaluation of segmentation results and for the optimization of segmentation algorithms.

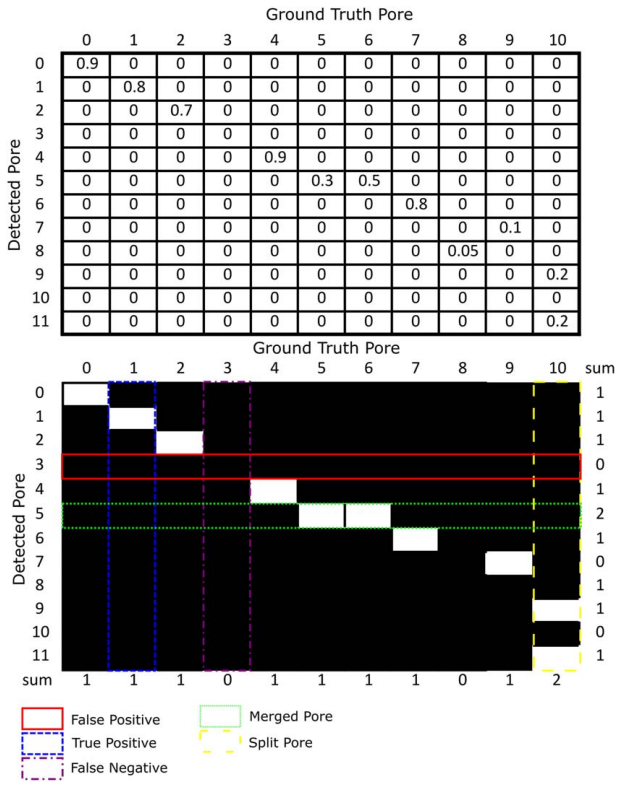

Fig. 7 IOU score matrix (top) and binary IOU score matrix (bottom). The cutoff for the local IOU score = 0.1. 11 GT and 12 segmented pores are considered.

$$\mathrm{IOU} = \frac{A \cap B}{A \cup B} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{1}$$

4.2 Instance Segmentation Evaluation. The instance segmentation evaluation task is to evaluate how well individual objects are segmented and detected. All GT pores with a diameter of 7 voxels or larger are near-spherical shape (perfect spheres that have been projected on a voxel mesh). Pore volumes and volume-equivalent spherical diameters (volume of a continuum sphere of the same volume) of the thresholded pores were found and compared to the volumes and volume-equivalent spherical diameters of the GT pores. We used the equivalent spherical diameter for ease of presentation and explanation of the concept. The volume-equivalent spherical diameter (D_{eq}) is the diameter of the sphere with equal volume to that of the pore (V_{pore}) and is given by Eq. (2). We also developed a volume-based metric, which can be applied to nonspherical pores. The aspect of object detection was also investigated through the POD framework.

$$D_{\mathrm{eq}} = \left(\frac{6 \cdot V_{\mathrm{pore}}}{\pi}\right)^{1/3} \tag{2}$$

When comparing individual thresholded pores to GT pores, it was not always straightforward to identify the correct pore pairs automatically. We identified the pore pairs based on IOU between individual GT pores and segmented pores, which is illustrated in Fig. 6(*b*). An IOU score matrix between all labeled GT pores and segmented pores and a binarized version (yes/no pore overlap) with the IOU score cutoff value = 0.1 are shown in Fig. 7. The user-defined cutoff value is used such that if the index is greater than or equal to that cutoff, there is sufficient pore overlap, while if the index is below the cutoff, insufficient pore overlap exists. In this simple example, 11 GT pores were considered, and 12 pores appeared after the segmentation process. Diagonal elements of the matrix tended to be nonzero due to the same labeling sequence as that applied to GT images especially if there was a low false positive rate. The binarized IOU score matrix enabled the automatic identification of TP, FP, and FN by summing over the columns and rows. The value of TN for an object detection task is not meaningful as it is related to the background. These confusion matrix components as shown in Fig. 6(*b*) are evaluating different tasks from those described for a semantic segmentation evaluation of Fig. 6(*a*). If the row sum was equal to 1, it was a TP—one GT pore and one

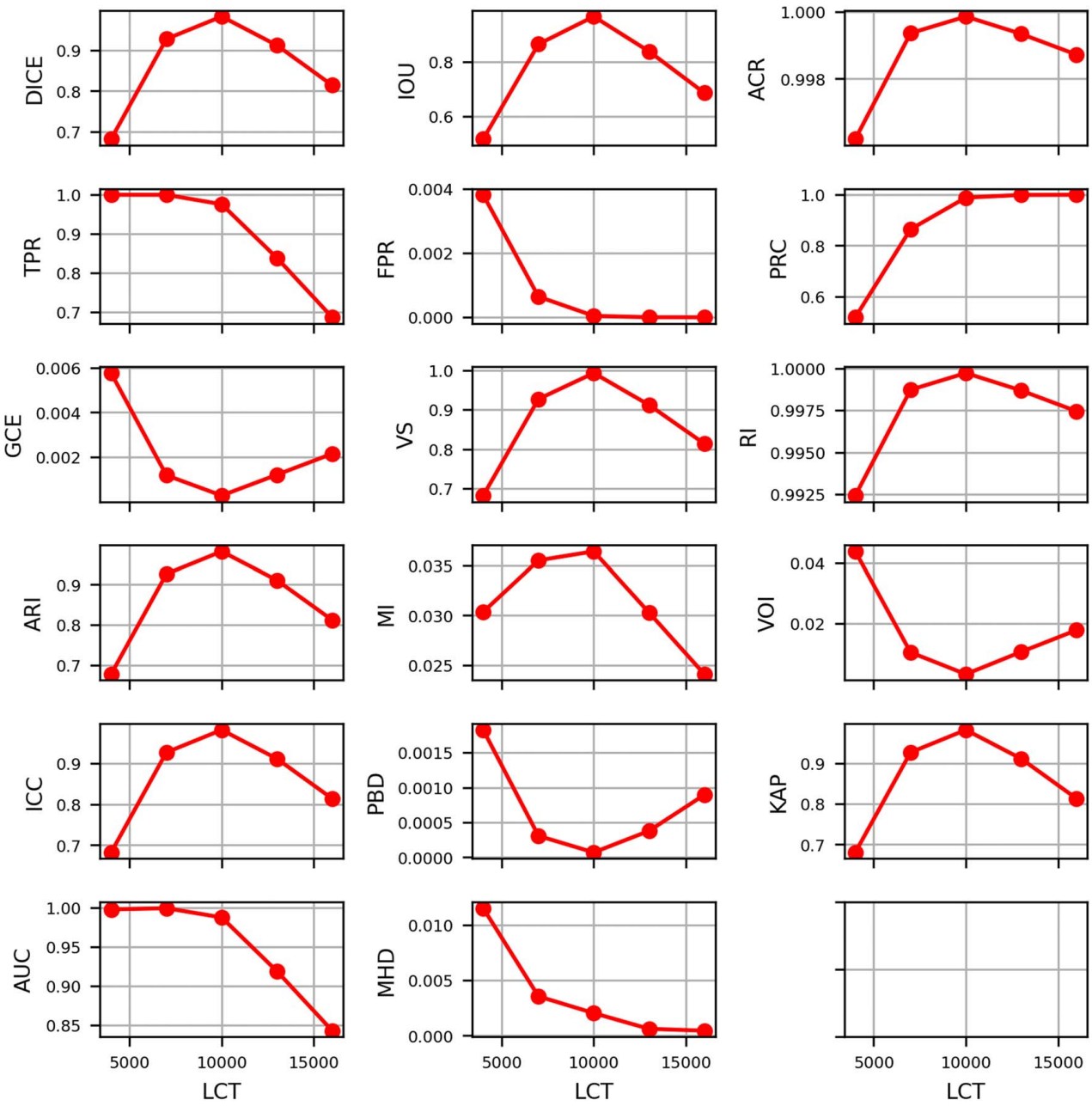

Fig. 8 Semantic segmentation evaluation results for different LCT values

overlapping segmented pore. If the column sum was equal to 0, it indicated FN—a single GT pore and no detection of an overlapping segmented pore. If the row sum was equal to 0, it indicated an FP. In addition to typical TP, FP, and FN, we identified situations of merged pores and split pores. If the row sum was larger than 1, it indicates a merged pore—two or more GT pores merged into one segmented pore. A merged pore may be found if two or more GT pores are positioned close to each other, and an example will be shown later. If the column sum was larger than 1, it indicated a split pore—a single GT pore split into two or more segmented pores. Figure 7 gives illustrations of these cases, marked with variously colored boxes.

5 Results

The various segmentations of the simulated XCT data using different local threshold parameters were evaluated using the GT image data and various evaluation metrics. This systematic comparison to GT provided the ability to optimize the local threshold parameters and to understand measurement uncertainty based on a particular algorithm parameter set. Optimization was carried out for each semantic segmentation metric, which, in general, produced somewhat different results.

5.1 Semantic Segmentation Evaluation. All unique semantic segmentation metrics were used to demonstrate the effect of a local thresholding algorithm parameter (LCT) for the commercial algorithm as shown in Fig. 8. The LCT was varied in five levels from 4000 to 16,000, while the LW was fixed at 11 voxels. The data set with 4 fpp was used for this example (SNR = 2×). For many evaluation metrics (DICE, IOU, ACR, GCE, VS, RI, ARI, MI, VOI, ICC, PBD, and KAP), clear patterns with maxima and minima were observed, which indicated the possibility of optimization. Some metrics such as ACR were not as sensitive to changes in the LCT parameter compared to metrics such as DICE or IOU due to imbalance of feature (positive) and background (negative) sizes.

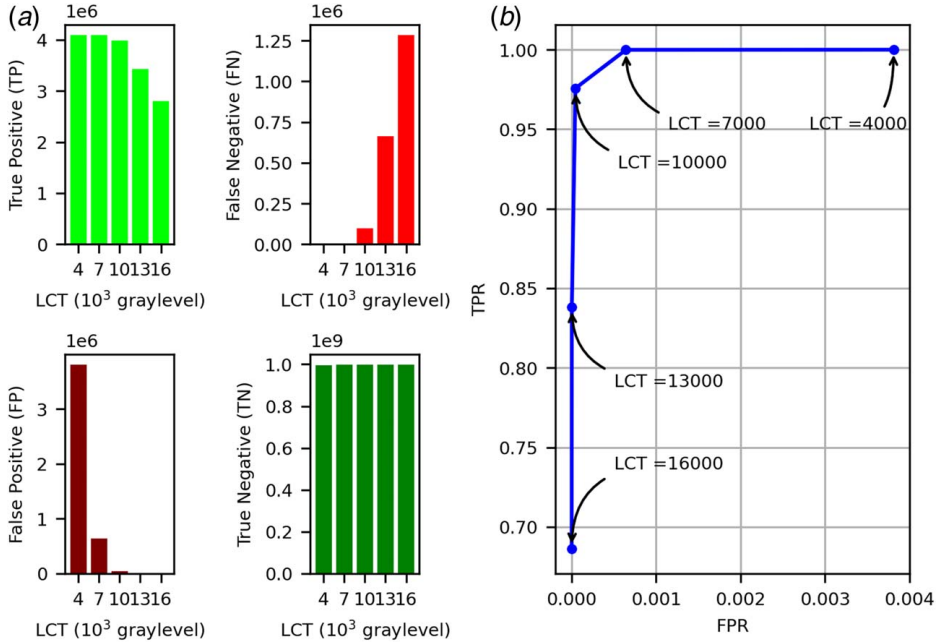

Fig. 9 (*a*) Confusion matrix variation for LCT and (*b*) receiver operating characteristic curve for varying LCT values

The background accounts for many more voxels than pores in this dataset, and therefore, the change in other phases is not significant. The confusion matrix (TP, FP, TN, and FN) in Fig. 9(*a*) also shows that the negative phase (i.e., FP + TN) is significantly larger than the positive phase, with TN counts dominating all others. The changes in the confusion matrix components are also shown in the figure. As the LCT increases, the FP and TP counts decreased, while FN and TN counts increased. The metrics such as DICE and IOU do not use the dominant TN counts; therefore, DICE or IOU present higher sensitivity to changes in positive phase (i.e., TP + FN) and are preferred over ACR. The TPR, FPR, and PRC do not show clear maxima nor minima. These metrics are usually combined as the ROC curve (i.e., TPR versus FPR) or as a precision-recall curve (i.e., PRC versus TPR) to provide more meaningful results. The ROC curve (i.e., TPR versus FPR) in Fig. 9(*b*) also shows the effect of LCT values. An optimized value of LCT may be found by minimizing FPR while maximizing TPR, which is near the knee point of the curve.

Comparison at the individual pore level shows some interesting similarities and differences. Segmentation results of a small pore (≈11 voxel diameter) and a large pore (≈31 voxel diameter) are shown in Fig. 10 for three levels of LCT using the commercial algorithm. The images are 2D cross-sectional images at the mid-height of the GT pores. The GT image of the large pore is shown in Fig. 10(*a*), and segmented images of the pores are shown in Figs. 10(*b*)–10(*d*) using different LCT values. The GT image of the small pore is shown in Fig. 10(*e*), and segmented images of the pores are shown in Figs. 10(*f*)–10(*h*) using the same LCT values. The LW size was fixed at 11 voxels for these examples. The global IOU scores of the entire 3D image, shown in Fig. 10(*i*) for different LCTs, display a maximum. The behavior of the IOU score for just the large pore (Fig. 10(*j*)) is very similar to that of the global IOU score, while the IOU score for the small pore (Fig. 10(*k*)) behaves differently as a function of LCT. The result shows that at LCT = 10,000, the pore measurement result is close to the GT for the large pore, but not for the small pore. The larger pores occupied more volume in the images compared to the smaller pores, and so the global IOU score and the IOU scores of the larger pores show a high correlation for the same LCT value. On the other hand, smaller pores present lower contrast (as shown in Fig. 4) and, when combined with the complex interaction of the edge blurring effect, the LCT value optimized for the larger pores may not be optimal for smaller pores.

The Bernsen algorithm shows a different behavior in Fig. 11. The LCT value was varied from 2500 to 65,000 (2500 to 30,000 shown in the figure) by increments of 2500, and three LW sizes were used: 5 voxels, 11 voxels, and 17 voxels. The SNR was also varied in three levels. No clear maxima or minima were observed for the variation of the local thresholding parameters within the bounds selected for the study. Instead, only a minimal change is observed after LCT became large enough, which meant between about 12,000 and 18,000 depending on SNR, LW, and the semantic segmentation evaluation metric. It is also possible to observe that the increase in LW size increases the LCT needed to reach the plateau. The effect of increasing the SNR value can be seen using the IOU score, which increased also with SNR. Similar to the results of the commercial algorithm, DICE and IOU show higher sensitivities to parameter changes compared to ACR because for any LCT, ACR is greater than 0.8, whereas DICE and IOU occupy the entire possible range of 0–1. Figure 12 shows thresholding results for the same small diameter (≈11-voxel diameter) pore and the large diameter (≈31-voxel diameter) pore shown for the commercial algorithm at different LCT values while keeping LW = 11 voxels. Details of the Bernsen thresholding algorithm are provided in the Supplemental Materials, but a few key steps are covered here to aid interpretation of the results. Based on a local contrast value and LCT, the algorithm first determines whether the local window is within a single-phase region or a dual-phase region. Then, the voxel of interest is thresholded to 0 or 1 depending on the voxel intensity. By using a low LCT, a higher number of FP voxels were found around the segmented pores. Once the LCT reaches a sufficiently high value, minimal changes were observed to thresholding results, which explains why the semantic segmentation values are almost constant in Fig. 11 for large LCT values.

The commercial local thresholding algorithm, the Bernsen local thresholding algorithm, and the Otsu global thresholding algorithm are compared in Fig. 13 for the IOU metric. As seen in Fig. 13(*a*), both local thresholding algorithms outperformed Otsu's method when algorithm parameters were suitably chosen. The maximum metric value of the commercial algorithm was higher than that for the Bernsen algorithm. The Bernsen algorithm, on the other hand, was more tolerant of inexact choices of the user-defined parameters

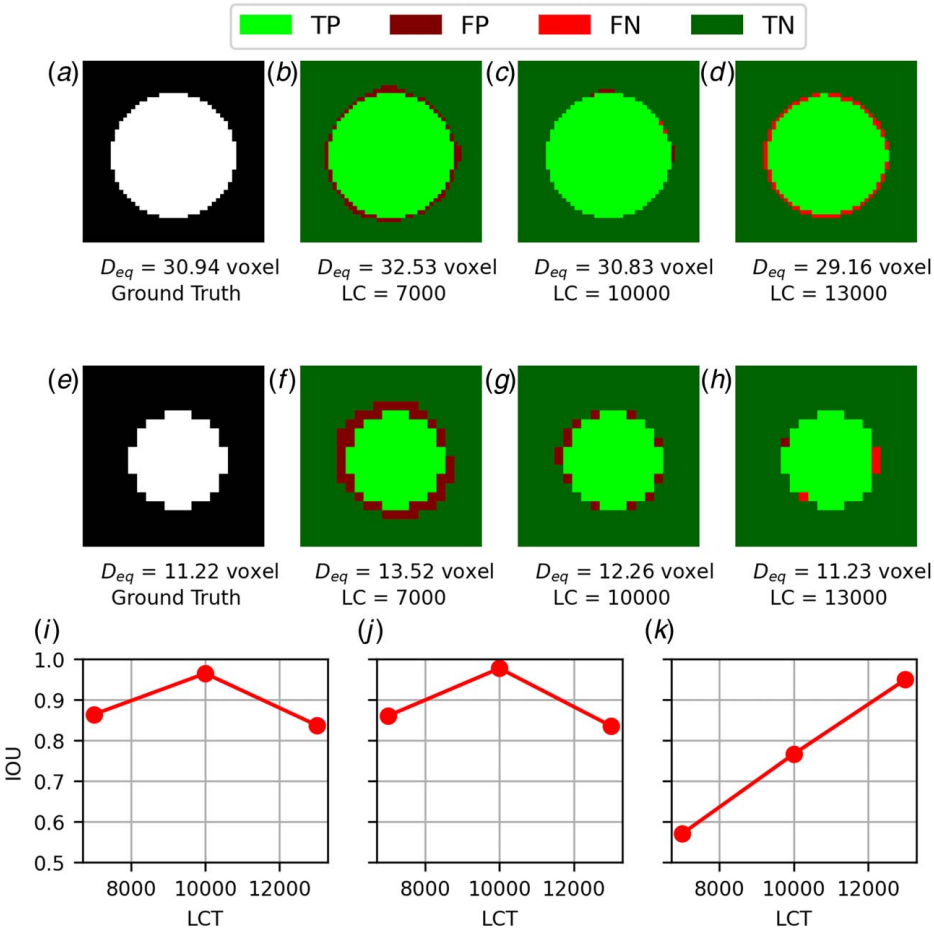

Fig. 10 (*a*)—(*d*) Thresholding results of a large pore (≈31 voxel diameter) at three LCT values, (*e*)—(*h*) thresholding results of a small pore (≈11 voxel diameter) at the same three LCT values, and IOU score at the same three LCT values for (*i*) the entire image, (*j*) the large pore, and (*k*) the small pore

(large flat regions for the Bernsen algorithm curves in Fig. 13(*a*)). Figure 13(*b*) compares confusion matrix components of the commercial algorithm (LCT = 10,000 and LW = 11 voxel), the Bernsen algorithm (LCT = 15,000 and LW = 11 voxel), and Otsu's algorithm. Using the confusion matrix components, the IOU scores were computed as shown in Table 3. The commercial algorithm performed better than the Bernsen local thresholding algorithm and the Otsu global thresholding algorithm at its best settings.

Many overlap-based metrics such as DICE and IOU account for differences in both the volume and the shape of the pores, and they provided good measures here. However, this type of metric can be problematic when segmentation shows misalignment or significant deviation in shape. In our data set, segmented pores were almost concentric while preserving near-spherical shape. Metrics such as ACR, however, are sensitive to the size of background (e.g., TN), and the metric may not be sensitive to changes in algorithm parameter values. VS only accounts for volume, and it may provide misleading information for more complex pore shapes, but it gave acceptable results for our data sets. MHD focuses on distances of cluster center rather than feature size or shape, and it may not be a good metric for evaluating segmentation accuracy of size or shape.

5.2 Optimization of Parameters. In this section, we demonstrate a method to optimize the local thresholding algorithm parameters, LCT and LW, for the commercial code. This is possible as there were distinct maxima or minima for most of the semantic segmentation evaluation metrics. A similar approach could be used for the Bernsen algorithm, but the goal would change from finding the optimal parameter values to finding values past which there is no further improvement. This could change some details, but the general method, consisting of the following three steps, would be the same: (1) vary the local thresholding parameters (and others of interest such as the SNR) according to an appropriate designed experiment (see, e.g., Ref. [54]) and record the evaluation metric for each parameter combination; (2) fit an appropriate model (e.g., a polynomial model) to the experiment results; and (3) use the fitted model to select the optimal local thresholding parameters. Using a designed experiment and model for interpolation would be more resource efficient than a brute force search over a large space of parameter combinations.

We varied image SNR, LCT, and LW size in the commercial local thresholding algorithm to observe how the best local thresholding parameters changed with the SNR. We constructed an optimally blocked response surface design [55] since different combinations of LCT and LW could be applied to the same simulated image, but the SNR is fixed for a given image. Four images (four blocks) were simulated. Three values of the SNR were used, ≈23.9, ≈50.3, and ≈77.1, and the SNR 50.3 was repeated once to obtain the fourth image. Nine combinations of the LCT and LW (some repeated) were then assigned to each image using the optBlock function of the AlgDesign package [56] for R [57]. The LCT values were 7000, 10,000, and 13,000, and the values of the LW size were 5, 11, and 17 voxels. A total of 36 experimental runs were implemented for the study.

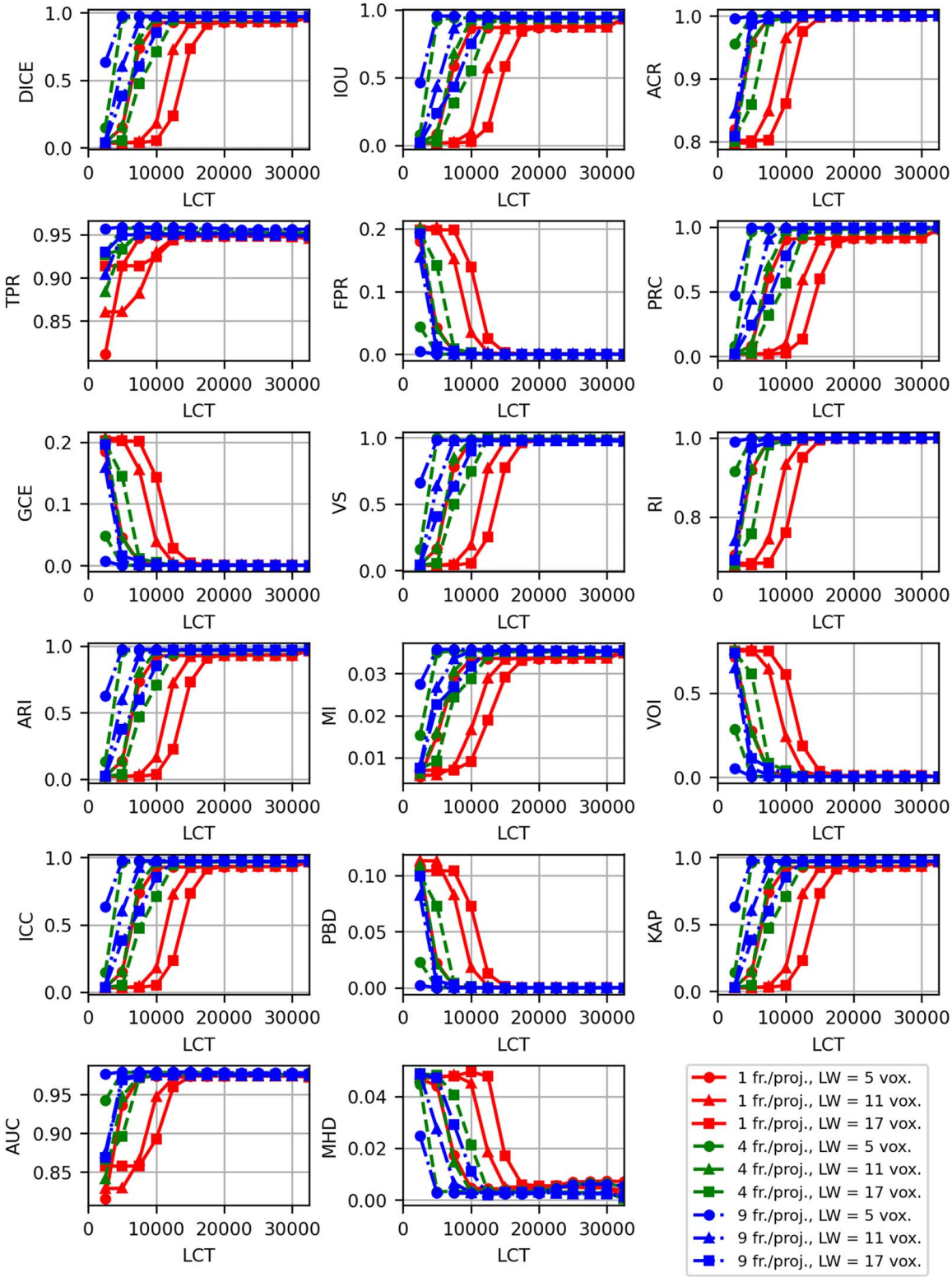

Fig. 11 Evaluation metric results of the Bernsen algorithm for varying parameters and SNRs

To find an optimal combination of LCT and LW for a given value of the SNR and a given metric, a full quadratic response surface was fitted where the response variable is the metric, and the predictor variables are the SNR, LCT, and LW. Since the experiment was designed and conducted as a split-plot experiment (see, for example, Chapter 9 of Ref. [54]) that structure is appropriately accounted for using the lmer function from the lme4 package [58] for R [57]. The fitted surface was used to find the optimal combination of LCT and LW, which can occur on the boundary of the experimental region defined by the rectangle $[7000, 13{,}000] \times [5, 17]$ (the extremes of LCT and LW considered). Confidence intervals were constructed using a parametric bootstrap algorithm [59]. In cases where a maximum was sought, relatively small values of the metric under study were removed so that they would not unduly

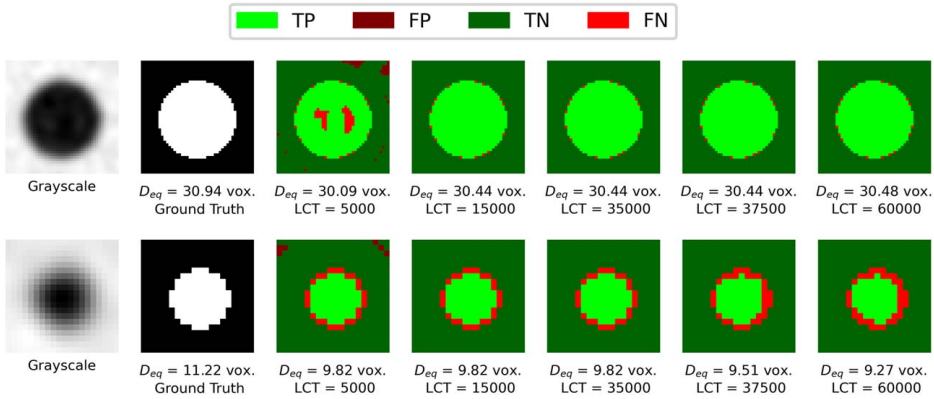

Fig. 12 Bernsen algorithm thresholding results for a large pore (≈31-voxel diameter, top row) and a small pore (≈11-voxel diameter, bottom row). LW = 11 voxel in this case.

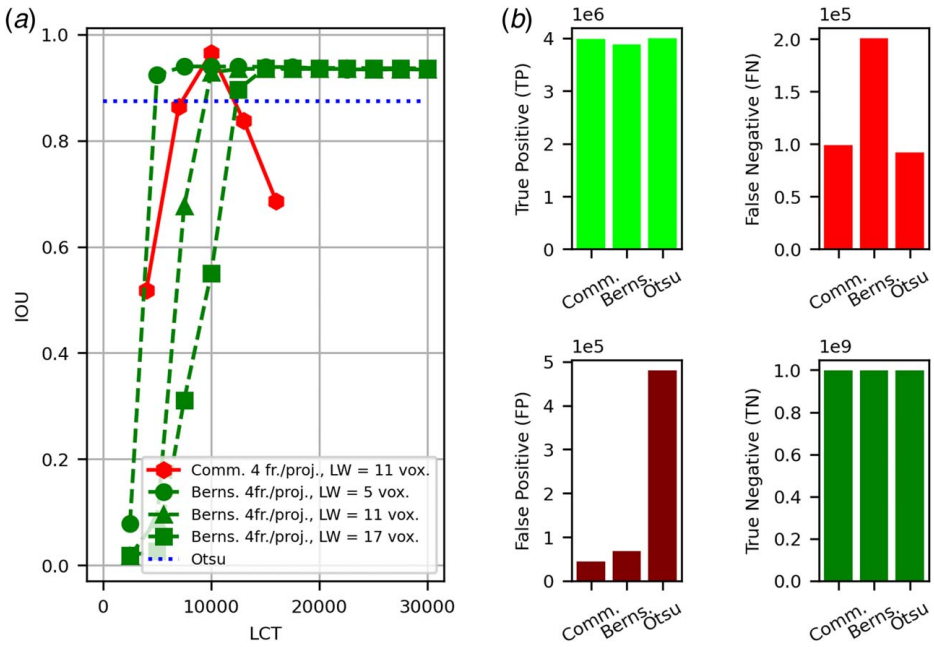

Fig. 13 (*a*) Comparison of commercial, Bernsen, and Otsu algorithms for IOU. (*b*) Comparison of confusion matrix components for commercial algorithm (LCT = 10,000 and LW = 11 voxel), Bernsen algorithm (LCT = 15,000 and LW = 11 voxel), and Otsu algorithm.

influence the fitted response surface. In cases where a minimum was sought, relatively large values were removed. The respective cutoffs for excluding an experimental run were the median value of the metric ±6× the median absolute deviation.

The algorithm segments an object smaller than GT for lower LCT while it segments the object larger than the GT for higher LCT. Hence, we observed clear maximum (or minimum) points for many semantic segmentation evaluation metrics. The presence of a maximum (or minimum) point for the metric away from the boundary of the experimental region is also related to the proper selection of parameter boundaries where we expect to observe maximum (or minimum) points. The optimized segmentation parameters are shown in Fig. 14. The segmentation software rounds up and down the optimized values to the nearest integer, and rounded values are shown for both the optimized values and their confidence intervals. The DICE and FMS metrics showed the same results as they are the same metrics. The FPR and TNR also showed the same results since they are monotonic transformations of each other as mentioned earlier. Several metrics produced the same optimized parameters, e.g., DICE, ARI, ICC, and KAP. The ACR and RI metrics did as well. The optimization results will be further discussed with pore size error plots in a later section. The optimized LCT value tends to reduce, and the LW value tends to increase as SNR improves. The mean-squared error (MSE) metric, the mean of the squared pore size errors, was also used as a metric for optimization and will be explained in the next section. The MSE metric is an evaluation metric for instance segmentation, but it was also used as a metric for optimization.

The total pore volume for each pore size is shown in Fig. 15(*a*). The IOU score contribution of each pore size is shown in Fig. 15(*b*). The IOU score was chosen as an example to show how performance metrics are often dominated by the large pores. These were determined by breaking the summation defining the IOU score into partial sums over the pore sizes. Intersection values for each pore size were divided by the same union values for all pores (same denominator for each numerator). The linear correlation between pore volume and pore size (Fig. 15(*c*)) shows that the dominant pore volumes also dominate this metric-based optimization.

Table 3 IOU score comparison of the commercial, Bernsen, and Otsu algorithms

	Commercial	Bernsen	Otsu
LCT (gray level)	10,000	15,000	N/A
LW (voxel)	11	11	N/A
IOU	0.965	0.935	0.874

5.3 Pore-level Segmentation Evaluation

5.3.1 Pore Sizing Error. Understanding the accuracy of individual pore size measurements is of great interest in many engineering applications. The individual pore sizing errors using the optimized local thresholding parameters were computed for the commercial algorithm, for hand-selected local thresholding parameters for the Bernsen algorithm, and for the Otsu algorithm. The segmentation results that were optimized for the IOU metric and applied to the 4 fpp data are used as an example to demonstrate the pore-level evaluation process. Based on the pore pairs found from the binary IOU score matrix (IOU cutoff value = 0) described in Sec. 4, an absolute pore diameter error ($E_{D_{eq}} = D_{th} - D_{gt}$) plot is shown in Fig. 16(*a*) based on the volume-equivalent spherical diameter of both GT pores (D_{gt}) and segmented pores (D_{th}). In this case, two merged pores were identified from the analysis corresponding to four outlier data points shown in Fig. 16(*a*). The merged pores result in outlier data points as the GT pore diameter was subtracted from the much larger merged pore diameter. These merged pores were already identified from the binary IOU score matrix. While it may be possible to separate such merged pores through a watershed separation algorithm [60], we did not consider such additional steps in order to focus solely on the segmentation algorithms. Figure 16(*b*) shows two GT pores, which had a distance of 3.1 voxels from surface to surface, and Fig. 16(*c*) shows how the two pores were merged in a narrow neck. The approximate surface-to-surface distance between the two neighboring GT pores was found by subtracting the sum of the two equivalent spherical radii from the distance between the spheres' center voxels.

In Fig. 16(*d*), the merged pores are removed from the absolute pore size error plot. For the evaluation task in this article, we do not consider merged pores or split pores as they can skew the average analysis result. Smaller-sized pores displayed larger average deviations from the GT and more variability in the individual deviations compared to larger pores. Standard deviations and 95% confidence

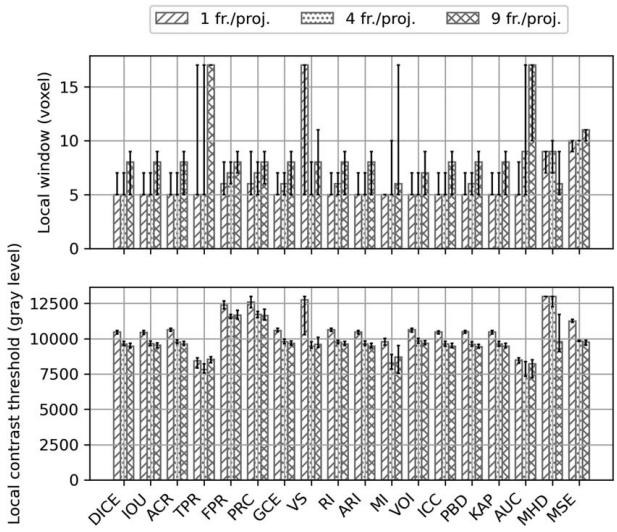

Fig. 14 Optimized LCT and LW parameters and uncertainty bounds

bounds for the mean (light gray band) of the pore size errors are shown in Fig. 16(*e*). The confidence bounds were based on the standard error of the mean (std/sqrt(n)) multiplied by the Student's t distribution percentile based on the target confidence (95%) and degrees-of-freedom (n–1) with the number of detected pore pairs (n). For example, for 11 voxels, std = 0.119, n = 20, Student's t distribution factor = 2.093, sqrt(n) = 4.47, and the confidence bound became very narrow (±0.0279). In this case, we found an average error of about 1 voxel for the 11-voxel diameter pore and an average error of approximately 0 voxel for 25-voxel diameter pore. The MSE, $(\frac{1}{N}\sum_{i=1}^{N}(D_{gt} - D_{th})^2)$, can be found for all pore sizes, where N is all segmented pore pairs with a single GT pore. The MSE can be used as an additional metric for optimization. If used in its aforementioned raw form, based on our observations from Fig. 16, MSE will prioritize smaller pores with larger pore size errors, in contrast to IOU, which we established already prioritizes large pores.

5.3.2 Probability of Detection. A binary POD analysis was also carried out on these results following the principles of

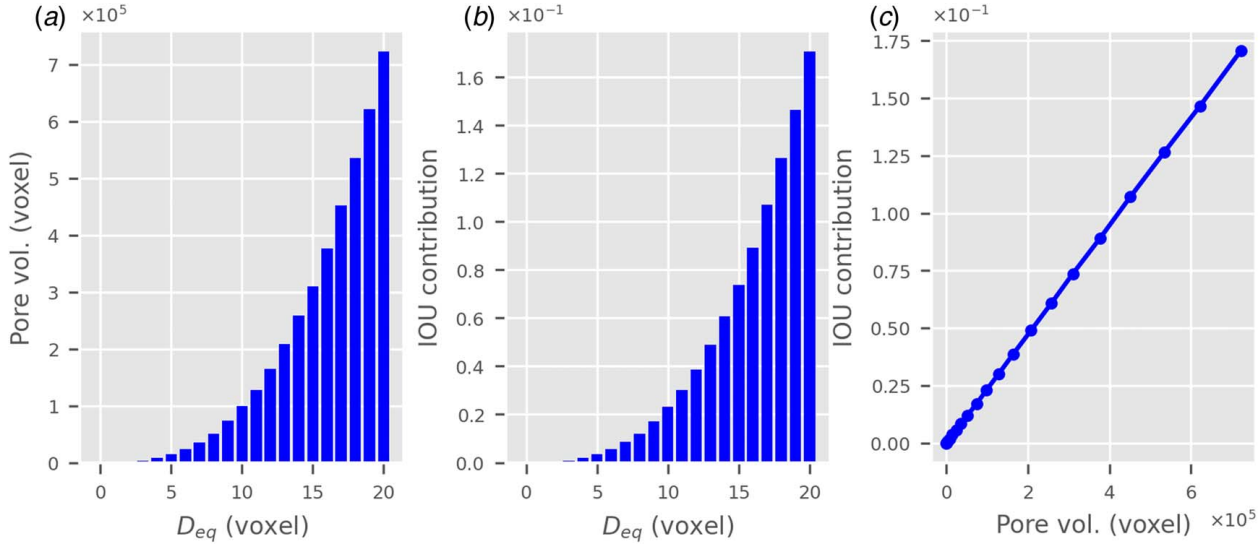

Fig. 15 (*a*) Pore volume distribution of the system, (*b*) IOU score contribution of individual pore sizes, and (c) correlation between pore volume and IOU score contribution

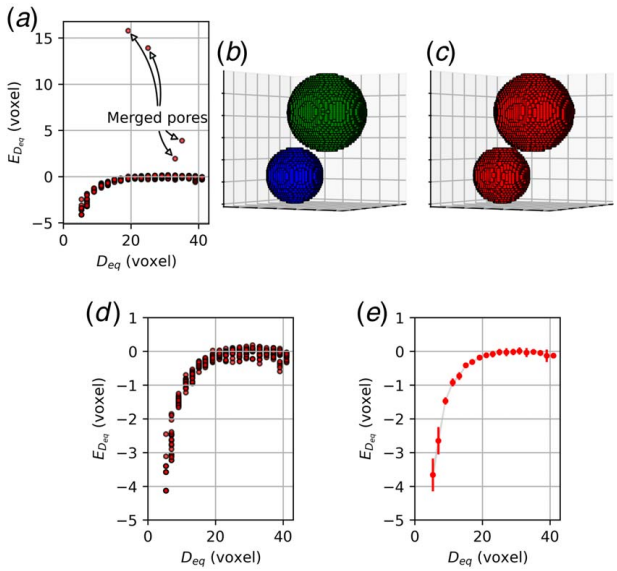

Fig. 16 (*a*) Pore size error ($E_{D_{eq}}$) plot showing outliers due to merged pores, (*b*) GT pores separated by approximately 3.1 voxels surface to surface, and (*c*) a merged pore after segmentation (*d*) pore size error plot with the merged pores removed, (*e*) pore size error plot showing mean (red point), ±1 standard deviation (red bar), and 95% confidence bound (gray band)

MIL-HNBK-1823A [61]. Previous XCT POD studies [8] focused on a signal-response ($\hat{a}$ versus a) POD analysis, which is an application-specific method that depends on a user-defined decision threshold when used for inspection. The segmented pore volume can be used as the signal ($\hat{a}$), and the variation in segmented pore volumes due to possible variations in the thresholding algorithm parameter was propagated into POD curve uncertainty in Ref. [17]. In contrast, for this binary POD analysis, a pore was considered to be detected if it was identified from the IOU score matrix and the chosen IOU score cutoff value. The detected pores are plotted as ones and undetected pores are plotted as zeros in Fig. 17 using the same data sets shown in Fig. 16. A pore was identified as found regardless of the sizing accuracy. A logistic regression model (see, e.g., Ref. [62]) with the GT diameter as the predictor and the zeros and ones, indicating nondetection and detection, respectively, as the response, was used to estimate the POD curve. Because there are a mix of zeros and ones only at the GT diameter of 5 voxels, the Bayesian statistical inference paradigm with weakly informative proper prior distributions was used to estimate the slope and intercept of the logistic regression model as well as quantify uncertainty. The same approach can be used when there

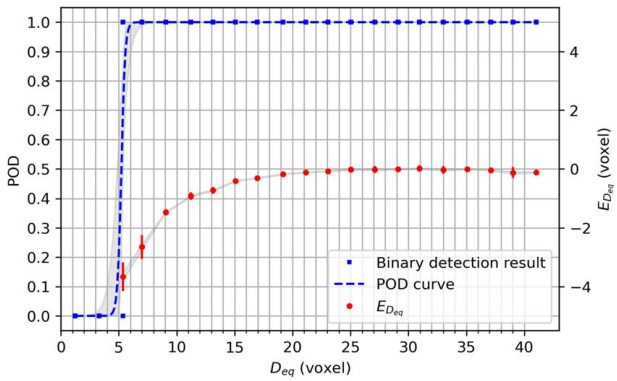

Fig. 17 Binary POD curves (left *y*-axis) and pore size error plot (right *y*-axis)

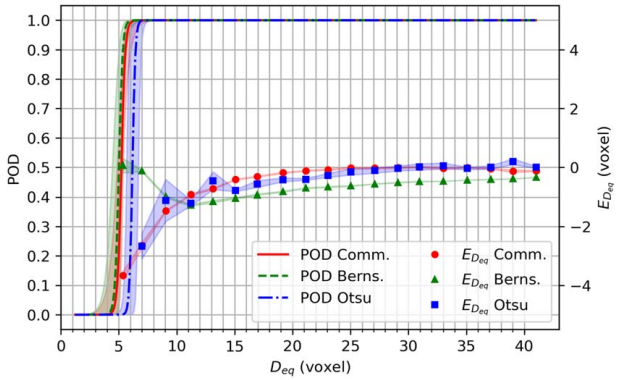

Fig. 18 Comparison of POD curves and pore diameter error plots for the commercial (LCT = 9623 and LW = 5 voxel), Bernsen (LCT = 15,000 and LW = 11 voxel), and Otsu algorithms

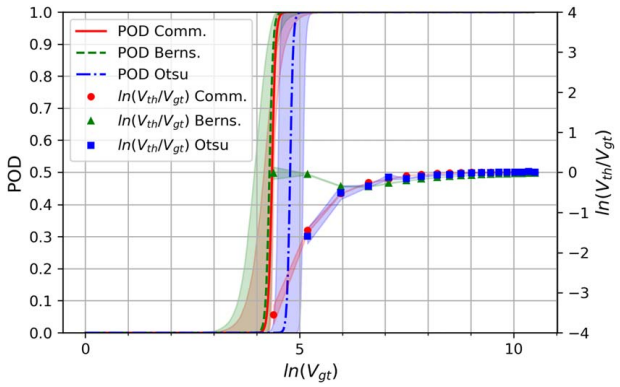

Fig. 19 POD curves and volume error plots in log scale for the commercial (LCT = 9623 and LW = 5 voxel), Bernsen (LCT = 15,000 and LW = 11 voxel), and Otsu algorithms

is complete separation, i.e., when no GT diameters have a mix of zeros and ones. The corresponding pore size errors are plotted in Fig. 17. In this case, we can expect to have an approximately 65% POD for a pore with a 5-voxel diameter, but with an average sizing error of about 3.5 voxels. The two metrics are complementary as high POD does not necessarily mean high accuracy in size measurements.

5.3.3 Comparison of Segmentation Algorithms. Figure 18 compares the three segmentation algorithms (commercial, Bernsen, and Otsu) for POD and pore size error. The commercial algorithm local thresholding parameters are optimized based on the IOU metric (LCT = 9653 and LW = 5 voxels). The corresponding parameters for the Bernsen algorithm are LCT = 15,000 and LW = 11 voxels. Using the optimized parameters, the commercial algorithm resulted in an IOU score of 0.975, which is a small improvement over the value of 0.965 found based on LCT = 10,000 and LW = 11 voxel. The commercial algorithm provided an average pore size error of about −1.5 voxels for the 9 voxel diameter pores while the POD was approximately 95%. Although the Bernsen algorithm showed slightly higher POD compared to other algorithms, it performed poorly for pore sizing accuracy. The Otsu algorithm showed the worst POD curve, but the pore size error for larger pores ($D \geq 15$ voxel) was comparable to the commercial algorithm. The variability in the individual pore size errors of smaller pores ($D < 15$ voxel) was the largest (in magnitude) for Otsu's algorithm.

In addition to evaluating pore diameter error, a pore volume-based metric was also proposed. In the case of a pore with a

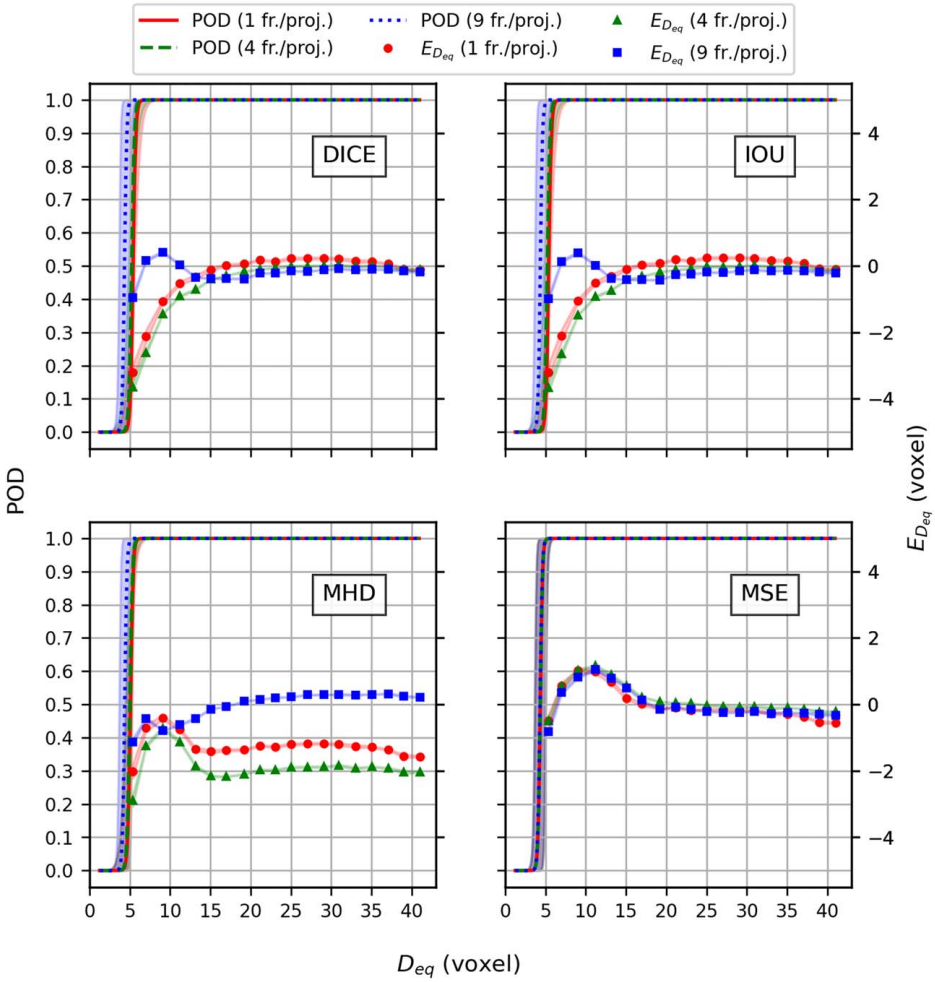

Fig. 20 POD curves (left *y*-axis) and pore size error plots (right *y*-axis) of DICE, IOU, MHD, and MSE

spherical shape, the use of D_{eq} is completely valid and equivalent to the use of volume. In many cases, however, we expect to evaluate nonspherical pores, and the evaluation of these pores based on a derived metric becomes problematic. As a result, we also proposed to use $\ln(V_{th}/V_{gt}) = (\ln V_{th} - \ln V_{gt})$ as a basis to understand segmentation error, where V_{th} and V_{gt} are pore volumes of the thresholded image and the GT image, respectively. Figure 19 shows example POD curves and volume error plots in log scale. Volume-based metrics were also used for previous XCT signal-response POD analyses [8,17].

5.4 Comparison of Optimization Results. Using the optimized local thresholding parameters (LCT and LW) presented in Fig. 14, pore size errors and POD curves are compared in Fig. 20. The metrics DICE, IOU, MHD, and MSE used for the optimization processes were selected for comparison to highlight their similarities and differences. Other results are also shown in the Supplemental Materials. The POD generally improves for increased SNR values, which means that noise reduction improved the detection of smaller pores. The selected local thresholding parameters using DICE and IOU are close to each other, and the pore size errors and POD curves are similar. The average errors tend to reduce in magnitude as the SNR value increased. Pores with diameters larger than about 15 voxels showed a small average pore size error. The 9 fpp data show an average pore size error pattern that is different from the equivalent 1 fpp and 4 fpp data. The average pore size error is within ±0.5 voxel for smaller pore sizes ($D_{eq} < 15$ voxel), while the average pore size error is close to 0 for larger pore sizes ($D_{eq} \geq 15$ voxel). Optimizing the local thresholding parameters based on the MHD metric leads to the largest pore size errors. This occurs because the MHD metric compares only the centers of the GT and segmented pores. The size of the pores is not considered in the computation. MSE-based optimization, on the other hand, shows similar pore size error trends for all SNRs. The MSE for all pores in an image, based on optimizing the local thresholding parameters for the selected metric and SNR

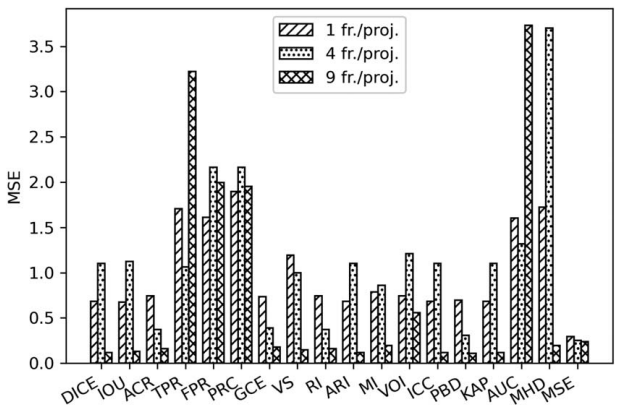

Fig. 21 MSE of pore diameter error for optimization results based on all metrics

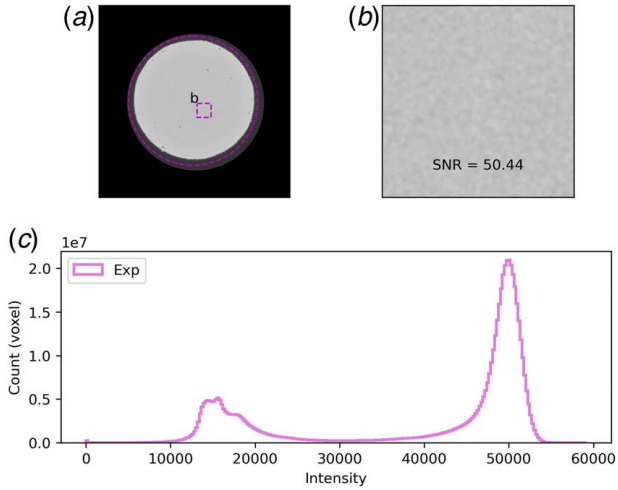

Fig. 22 (*a*) Example XCT slice, (*b*) ROI, and (*c*) histogram of the experimental data

combination, are shown in Fig. 21. MSE averages over all pore sizes, so the metric may be more sensitive to the larger errors at smaller pore sizes. The MSE value generally decreased as the SNR value increased. However, in some cases, the MSE values for the 4 fpp and/or 9 fpp data are higher than for the 1 fpp data. MSE-based optimization showed low MSE values for all SNRs, which is reasonable since the optimization was carried out to minimize MSE for each value of SNR. The expected trend of decreasing MSE with increasing SNR is also observed for the MSE-based optimization.

6 Discussion

6.1 Calibration of Experimental Pore Size Measurement. Starting with clear GT information as the input to simulated XCT data provided clear benchmark information that enabled evaluation and comparison of XCT flaw detection algorithms. In addition, there is the possibility of using the simulated data as the basis for choosing detection algorithm parameters and estimating pore size variability and bias in XCT experimental data. VDI/VDE 2630 part 2.1 [23] mentions a few essential metrological characteristics such as the component shape, maximum transmission lengths, and the linear attenuation coefficient for meeting similarity conditions. The guideline does not provide details on how closely these conditions should match, and there may be a valid regime of the XCT simulation models for users to carefully consider. In this case, the XCT simulation was carried out based on experimental data with a similar part dimension, shape, and material as described in Ref. [24], which was an additively manufactured nickel superalloy cylinder (3 mm diameter and 3 mm height). All simulation parameters in this work were identically set to match image quality parameters of the experiments. There are differences in the pore size distribution and the reconstruction algorithm as the vendor software did not allow reconstruction of projection data acquired by different systems or from simulation software. The SNR value of the experimental data (50.44) is similar to that of the simulated data (50.83) with 4 fpp. Since the sample porosity was relatively low, the differences in the pore size distributions were not expected to significantly affect the thresholding parameters, and similar SNR values were expected to satisfy the similarity conditions. Depending on the application, it may be possible to transfer the optimized thresholding parameters determined from the simulated data to experimental data. Figure 22(*a*) shows an example reconstructed slice of experimental data, and Fig. 22(*b*) shows an ROI with an SNR value of 50.44 within the ROI. Figure 22(*c*) shows the grayscale histogram of the experimental data, which has peaks at similar locations as that of the XCT simulation (shown in Fig. 3(*g*)).

The segmented pore sizes are not always accurate, but the pore size measurement can be calibrated based on the simulated data. An $\hat{a}$ versus a curve used for a signal-response POD analysis is plotted from the simulated data in Fig. 23(*a*), where a refers to the GT pore sizes and $\hat{a}$ refers to corresponding signal (in this case, it is the segmented pore sizes). A seventh-degree polynomial curve was fit to the data, which well describes the variation in $\hat{a}$ as a function of a. Our interest was to predict a values and estimate uncertainty in those predictions based on observed values of $\hat{a}$, and therefore, an inverted curve is found in Fig. 23(*b*). Two example pores and segmentation results are shown in Figs. 23(*c*) and 23(*d*). The pore in Fig. 23(*c*) has a measured volume of 3675 voxels with D_{eq} of 19.14 voxels. Based on the calibration curve, the pore can be calibrated to 19.32 ± 0.01 voxels. The pore in Fig. 23(*b*) has a segmented volume of 391 voxels with a D_{eq} of 9.07 voxels, which is calibrated to be 10.22 ± 0.02 voxels. A higher level of nonlinearity was observed for the smaller pore sizes, and the difference between measured and calibrated values is higher in this range.

6.2 Considerations for Developing X-Ray Computed Tomography Flaw Data. NIST is planning to develop a series of XCT data sets with flaws and associated GT data. We considered

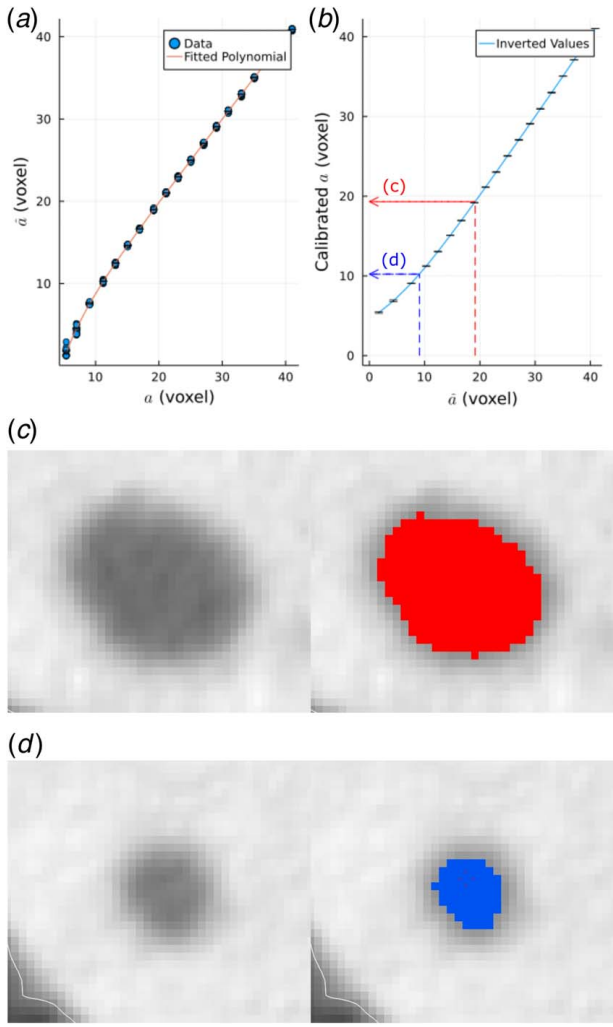

Fig. 23 (*a*) $\hat{a}$ versus a curve, (*b*) calibration curve, and (*c* and *d*) example segmentation results (left: grayscale image and right: original segmentation results)

spherical pores in this article for simplicity of presentation of the concept. We plan to include more complex pore shapes and orientations in the future datasets. Pore merging due to the close proximity of multiple pores may be further studied systematically. Pore splitting is expected to occur on more complex-shaped pores, especially at narrow necks. There are many possible variations in pore shapes and orientation. Experiment design can help efficiently explore such a large space. More complex part geometries must also be considered to account for imaging artifacts in the pore detection processes. Complex part shapes are more prone to imaging artifacts (e.g., beam hardening), which will make the defect detection process more challenging. The part surface determination step can also affect thresholding results if any pore is located very close to the part surface. The surface determination capabilities and implemented preprocessing steps can affect performance of detection algorithms. Defects showing lower intensity than the solid material were considered in this article, but higher intensity inclusions are occasionally observed in AM parts due to contamination. Detection algorithms should also find these types of inclusions.

A detection task may be carried out by a computer, a human observer, or a combination of the two. We focused on computer-only algorithm-based detection, but other types of detection tasks will be considered in the future and data sets will be prepared accordingly. For the evaluation of human observer-based detection, continuum-shaped pores can be used including sub-voxel sized flaws. Without evaluating the pore sizes, a binary POD analysis, similar to the one described in this article, will be the best method of evaluation. A sequential experiment design to more accurately characterize the transition region from nondetection to detection may be necessary. It is important to have both nondetections and detections at multiple GT pore sizes in the transition region.

7 Summary and Conclusions

We presented a full workflow of generating synthetic XCT phantoms with clear GT pores and evaluating XCT flaw detection algorithms with various metrics. This type of data set was useful for determining detection limits and flaw size measurement accuracy and uncertainty using a computer-based thresholding and detection algorithm. The workflow can be easily applied to various part designs and segmentation tasks. The simulated data avoided any manual labeling process for GT images. Using voxelated pores for XCT simulation eliminated uncertainty due to the discretization of a continuum mesh to generate GT images. To demonstrate evaluation processes, two different local thresholding algorithms (commercial and Bernsen) and a well-known global thresholding algorithm (Otsu) were compared and evaluated using many semantic segmentation evaluation metrics. The details of the commercial local thresholding algorithm are not disclosed to users, but it is possible to observe that the two local thresholding algorithms operate differently in principle. The commercial algorithm often showed clear maxima/minima when plotting semantic segmentation evaluation metrics against the corresponding value of the local contrast threshold, but the Bernsen algorithm showed different patterns. The semantic segmentation evaluation metrics provided a relatively quick and easy guide for comparing different algorithms. Each metric emphasizes different properties of segmentation, and the choice of metric depends on the intended application. For applications of evaluating structural integrity, accuracies in pore volumes, shapes, and positions would be of major interest. In this case, segmentation quality was good where most of the pores were segmented properly, and many evaluation metrics showed reasonable results. While the evaluation results depend on the algorithm parameters, the commercial algorithm showed the potential to perform better than the Bernsen algorithm. On the other hand, the Bernsen algorithm was more tolerant of user choices over a wide range of user-defined input parameters. Some semantic segmentation evaluation metrics may not be sensitive to smaller pores, which tend to occupy a smaller volume fraction of the images. Therefore, pore-level evaluations (pore sizing errors and binary POD curves) were also carried out to complement the semantic segmentation evaluations. The pore size error plots provided the level of measurement error, and the POD curves estimate pore detectability without considering the pore size measurement accuracy. The bias and variability in errors generally increased for smaller pores below approximately 15 voxels in diameter. Pores below 5 voxels in diameter were never segmented. We also presented a resource-efficient approach to optimize the thresholding algorithm parameters by leveraging the design of experiments. The optimization approach can be automated in principle, and it may be adopted by software vendors for applicable algorithms. We also introduced mean-squared pore size error as an additional metric for evaluation and optimization. When it was used as a metric for optimizing algorithm parameters, the derived MSE values after the optimization were found to be consistently low. The simulated data also provided a calibration curve to correct segmentation results and estimate uncertainty. We further discussed the possibility of transferring thresholding parameters and uncertainties determined from the simulated data to experimental data with similar part shapes/dimensions and SNRs. We demonstrated a method to calibrate experimental pore size measurements based on simulated data. NIST is planning to develop and distribute a series of datasets following similar methods described in this article as part of a program in XCT flaw detection challenges. The effect of complex part geometry and different reconstruction algorithms will be further investigated. These types of data and metrics will also be useful for the training and validation of machine learning-based flaw detection algorithms.

Conflict of Interest

There are no conflicts of interest.

Data Availability Statement

The data and information that support the findings of this article are freely available at: https://doi.org/10.18434/mds2-2846

References

[1] Todorov, E., Spencer, R., Gleeson, S., Jamshidinia, M., and Kelly, S. M., 2014, *America Makes: National Additive Manufacturing Innovation Institute (NAMII) Project 1: Nondestructive Evaluation (NDE) of Complex Metallic Additive Manufactured (AM) Structures*, Air Force Research Laboratory, Wright-Patterson Air Force Base, OH.

[2] Kim, F. H., and Moylan, S. P., 2018, "Literature Review of Metal Additive Manufacturing Defects," NIST Advanced Manufacturing Series (NIST AMS), National Institute of Standards and Technology, Gaithersburg, MD, pp. 100–116.

[3] Ng, G. K. L., Jarfors, A. E. W., Bi, G., and Zheng, H. Y., 2009, "Porosity Formation and Gas Bubble Retention in Laser Metal Deposition," Appl. Phys. A, **97**(3), pp. 641–649.

[4] Vandenbroucke, B., and Kruth, J. P., 2007, "Selective Laser Melting of Biocompatible Metals for Rapid Manufacturing of Medical Parts," Rapid Prototyp. J., **13**(4), pp. 196–203.

[5] King, W. E., Barth, H. D., Castillo, V. M., Gallegos, G. F., Gibbs, J. W., Hahn, D. E., Kamath, C., and Rubenchik, A. M., 2014, "Observation of Keyhole-Mode Laser Melting in Laser Powder-Bed Fusion Additive Manufacturing," J. Mater. Process. Technol., **214**(12), pp. 2915–2925.

[6] Beretta, S., and Romano, S., 2017, "A Comparison of Fatigue Strength Sensitivity to Defects for Materials Manufactured by AM or Traditional Processes," Int. J. Fatigue, **94**(2), pp. 178–191.

[7] Kim, F. H., Moylan, S. P., Garboczi, E. J., and Slotwinski, J. A., 2017, "Investigation of Pore Structure in Cobalt Chrome Additively Manufactured Parts Using X-Ray Computed Tomography and Three-Dimensional Image Analysis," Addit. Manuf., **17**, pp. 23–38.

[8] Kim, F. H., Pintar, A. L., Moylan, S. P., and Garboczi, E. J., 2019, "The Influence of X-Ray Computed Tomography Acquisition Parameters on Image Quality and Probability of Detection of Additive Manufacturing Defects," ASME J. Manuf. Sci. Eng., **141**(11), p. 111002.

[9] Sezgin, M., and Sankur, B. l., 2004, "Survey Over Image Thresholding Techniques and Quantitative Performance Evaluation," ELECTIM, **13**(1), pp. 146–168.

[10] Wong, V. W. H., Ferguson, M., Law, K. H., Lee, Y.-T. T., and Witherell, P., 2022, "Segmentation of Additive Manufacturing Defects Using U-Net," ASME J. Comput. Inf. Sci. Eng., **22**(3), p. 031005.
[11] Chisena, R. S., Engstrom, S. M., and Shih, A. J., 2020, "Automated Thresholding Method for the Computed Tomography Inspection of the Internal Composition of Parts Fabricated Using Additive Manufacturing," Addit. Manuf., **33**, p. 101185.
[12] Langner, O., Karolczak, M., Rattmann, G., and Kalender, W. A., 2009, "Bar and Point Test Patterns Generated by Dry-Etching for Measurement of High Spatial Resolution in Micro-CT," World Congress on Medical Physics and Biomedical Engineering, Munich, Germany, Sept. 7–12.
[13] ASTM E1695-20, 2020, *Standard Test Method for Measurement of Computed Tomography (CT) System Performance*, ASTM International, West Conshohocken, PA.
[14] Shepp, L. A., and Logan, B. F., 1974, "The Fourier Reconstruction of a Head Section," IEEE Trans. Nucl. Sci., **21**(3), pp. 21–43.
[15] Ching, D. J., and Gursoy, D., 2017, "XDesign: An Open-Source Software Package for Designing X-Ray Imaging Phantoms and Experiments," J. Synchrotron Radiat., **24**(2), pp. 537–544.
[16] Kazantsev, D., Pickalov, V., Nagella, S., Pasca, E., and Withers, P. J., 2018, "TomoPhantom, a Software Package to Generate 2D–4D Analytical Phantoms for CT Image Reconstruction Algorithm Benchmarks," SoftwareX, **7**, pp. 150–155.
[17] Kim, F. H., Pintar, A., Obaton, A.-F., Fox, J., Tarr, J., and Donmez, A., 2021, "Merging Experiments and Computer Simulations in X-Ray Computed Tomography Probability of Detection Analysis of Additive Manufacturing Flaws," NDT & E Int., **119**, p. 102416.
[18] Slotwinski, J. A., Garboczi, E. J., and Hebenstreit, K. M., 2014, "Porosity Measurements and Analysis for Metal Additive Manufacturing Process Control," J. Res. Natl. Inst. Stand. Technol., **119**, pp. 494–528.
[19] Anderson, T. L., 2017, *Fracture Mechanics: Fundamentals and Applications*, CRC Press, Boca Raton, FL.
[20] Fuchs, P., Kröger, T., and Garbe, C. S., 2021, "Defect Detection in CT Scans of Cast Aluminum Parts: A Machine Vision Perspective," Neurocomputing, **453**, pp. 85–96.
[21] Bircher, B. A., Wyss, S., Gage, D., Küng, A., Körner, C., and Meli, F., 2021, "High-Resolution X-Ray Computed Tomography for Additive Manufacturing: Towards Traceable Porosity Defect Measurements Using Digital Twins," Joint Special Interest Group Meeting Between EUSPEN and ASPE Advancing Precision in Additive Manufacturing, Virtual, Sept. 21–23.
[22] Rentala, V. K., Kanzler, D., and Fuchs, P., 2022, "POD Evaluation: The Key Performance Indicator for NDE 4.0," J. Nondestruct. Eval., **41**(1), p. 20.
[23] Bellon, C., Deresch, A., Gollwitzer, C., and Jaenisch, G.-R., 2012, "Radiographic Simulator aRTist: Version 2," 18th World Conference on Nondestructive Testing, Durban, South Africa, Apr. 16–20.
[24] Kim, F. H., Yeung, H., and Garboczi, E. J., 2021, "Characterizing the Effects of Laser Control in Laser Powder Bed Fusion on Near-Surface Pore Formation via Combined Analysis of In-Situ Melt Pool Monitoring and X-Ray Computed Tomography," Addit. Manuf., **48**, p. 102372.
[25] Kim, F. H., Pintar, A. L., Scott, J.-H. J., and Garboczi, E. J., 2023, "Simulated X-Ray Computed Tomography (XCT) and Ground Truth Images of Cylindrical Sample with Randomly Distributed Spherical Pores," National Institute of Standards and Technology, Gaithersburg, MD.
[26] Brauer, C., Aukes, D., Brauer, J., and Jeffries, C., 2020, VoxelFuse, https://github.com/cdbrauer/VoxelFuse
[27] Cignoni, P., Callieri, M., Corsini, M., Dellepiane, M., Ganovelli, F., and Ranzuglia, G., 2008, "MeshLab: An Open-Source Mesh Processing Tool," Eurographics Italian Chapter Conference, V. Scarano, R. De Chiara, and U. Erra, eds., Salerno, Italy, July 2–4.
[28] Jaenisch, G.-R., Bellon, C., Samadurau, U., Zhukovskiy, M., and Podoliako, S., 2006, "McRay—A Monte Carlo Model Coupled to CAD for Radiation Techniques," 9th European Conference on NDT, Berlin, Germany, Sept. 25–29.
[29] van Aarle, W., Palenstijn, W. J., Cant, J., Janssens, E., Bleichrodt, F., Dabravolski, A., De Beenhouwer, J., Joost Batenburg, K., and Sijbers, J., 2016, "Fast and Flexible X-Ray Tomography Using the ASTRA Toolbox," Opt. Express, **24**(22), pp. 25129–25147.
[30] van Aarle, W., Palenstijn, W. J., De Beenhouwer, J., Altantzis, T., Bals, S., Batenburg, K. J., and Sijbers, J., 2015, "The ASTRA Toolbox: A Platform for Advanced Algorithm Development in Electron Tomography," Ultramicroscopy, **157**, pp. 35–47.
[31] Feldkamp, L. A., Davis, L. C., and Kress, J. W., 1984, "Practical Cone-Beam Algorithm," J. Opt. Soc. Am. A, **1**(6), pp. 612–619.
[32] Bernsen, J., 1986, "Dynamic Thresholding of Grey-Level Images," International Conference on Pattern Recognition, Paris, France, Oct. 27–31, pp. 1251–1255.
[33] Otsu, N., 1979, "A Threshold Selection Method From Gray-Level Histograms," IEEE Trans. Syst. Man, Cybern., **9**(1), pp. 62–66.
[34] Gatos, B., Pratikakis, I., and Perantonis, S. J., 2006, "Adaptive Degraded Document Image Binarization," Pattern Recognit. **39**(3), pp. 317–327.
[35] Dice, L. R., 1945, "Measures of the Amount of Ecologic Association Between Species," Ecology, **26**(3), pp. 297–302.
[36] Jaccard, P., 1912, "The Distribution of the Flora in the Alpine Zone.1," New Phytol., **11**(2), pp. 37–50.
[37] Chinchor, N., 1992, "MUC-4 Evaluation Metrics," Proceedings of the 4th Conference on Message Understanding, Association for Computational Linguistics, McLean, VA, June 16–18, pp. 22–29.
[38] Martin, D., Fowlkes, C., Tal, D., and Malik, J., 2001, "A Database of Human Segmented Natural Images and Its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics," Proceedings Eighth IEEE International Conference on Computer Vision, Vancouver, British Columbia, Canada, July 7–14, Vol. 412, ICCV, pp. 416–423.
[39] Cárdenes, R., de Luis-García, R., and Bach-Cuadra, M., 2009, "A Multidimensional Segmentation Evaluation for Medical Image Data," Comput. Methods Programs Biomed., **96**(2), pp. 108–124.
[40] Reddy, R. A., Prasad, E. V., and Reddy, L. S. S., 2012, "Abnormality Detection of Brain MR Image Segmentation Using Iterative Conditional Mode Algorithm," Int. J. Appl. Inf. Syst., **5**(2), pp. 56–65.
[41] RamaswamyReddy, A., Prasad, E. V., and Reddy, L. S. S., 2013, "Comparative Analysis of Brain Tumor Detection Using Different Segmentation Techniques," Int. J. Comput. Appl., **82**(14), pp. 14–28. .
[42] Vadaparthi, N., Penumatsa, S. V., Yarramalle, S., and Murthy, P., 2011, "Segmentation of Brain MR Images Based on Finite Skew Gaussian Mixture Model With Fuzzy C-Means Clustering and EM Algorithm," Int. J. Comput. Appl., **975**(10), p. 8887.
[43] Rand, W. M., 1971, "Objective Criteria for the Evaluation of Clustering Methods," J. Am. Stat. Assoc., **66**(336), pp. 846–850.
[44] Hubert, L., and Arabie, P., 1985, "Comparing Partitions," J. Classif., **2**(1), pp. 193–218.
[45] Viola, P., and Wells Iii, W. M., 1997, "Alignment by Maximization of Mutual Information," Int. J. Comput. Vis., **24**(2), pp. 137–154.
[46] Russakoff, D. B., Tomasi, C., Rohlfing, T., and Maurer, C. R., 2004, *Image Similarity Using Mutual Information of Regions*, Springer, Berlin/Heidelberg, pp. 596–607.
[47] Meilă, M., 2003, *Comparing Clusterings by the Variation of Information*, Springer, Berlin/Heidelberg, pp. 173–187.
[48] Shrout, P. E., and Fleiss, J. L., 1979, "Intraclass Correlations: Uses in Assessing Rater Reliability," Psychol. Bull., **86**(2), pp. 420–428.
[49] Gerig, G., Jomier, M., and Chakos, M., 2001, *Valmet: A New Validation Tool for Assessing and Improving 3D Object Segmentation*, Springer, Berlin/Heidelberg, pp. 516–523.
[50] Cohen, J., 1960, "A Coefficient of Agreement for Nominal Scales," Educ. Psychol. Meas., **20**(1), pp. 37–46.
[51] Powers, D. M., 2010, "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation," Int. J. Mach. Learn. Technol., **2**(1), pp. 37–63.
[52] McLachlan, G. J., 1999, "Mahalanobis Distance," Resonance, **4**(6), pp. 20–26.
[53] Taha, A. A., and Hanbury, A., 2015, "Metrics for Evaluating 3D Medical Image Segmentation: Analysis, Selection, and Tool," BMC Med. Imaging, **15**(1), p. 29.
[54] Box, G. E. P., Hunter, J. S., and Hunter, W. G., 2005, *Statistics for Experimenters*, Wiley-Interscience, New York.
[55] Cook, R. D., and Nachtsheim, C. J., 1989, "Computer-Aided Blocking of Factorial and Response-Surface Designs," Technometrics, **31**(3), pp. 339–346.
[56] Wheeler, B., 2004, optBlock. AlgDesign. The R Project for Statistical Computing. https://www.r-project.org/
[57] R Core Team, 2022, *A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
[58] Bates, D., Mächler, M., Bolker, B., and Walker, S., 2015, "Fitting Linear Mixed-Effects Models Using lme4," J. Stat. Softw., **67**(1), pp. 1–48.
[59] Efron, B., and Tibshirani, R. J., 1994, *An Introduction to the Bootstrap*, Chapman and Hall/CRC, London .
[60] Kim, F. H., Penumadu, D., Gregor, J., Kardjilov, N., and Manke, I., 2013, "High-Resolution Neutron and X-Ray Imaging of Granular Materials," J. Geotech. Geoenvironmental Eng., **139**(5), pp. 715–723.
[61] Department of Defense, 2009, "Nondestructive Evaluation System Reliability Assessment," MIL-HNBK-1823A, Department of Defense.
[62] Myers, R. H., Montgomery, D. C., Vining, G. G., and Robinson, T. J., 2012, *Generalized Linear Models: With Applications in Engineering and the Sciences*, John Wiley & Sons, Hoboken, NJ.