SB^{3}C2021
Summer Biomechanics, Bioengineering and Biotransport Conference
June 14-18, Virtual

MACHINE LEARNING SEGMENTATION FOR DENSITOMETRIC PHANTOM CT DATA

Ashley N. Pernsteiner (1), Carla Winsor (1), Heidi-Lynn Ploeg (2), Corinne R. Henak (1)

(1) Department of Mechanical Engineering
University of Wisconsin - Madison
Madison, Wisconsin, USA

(2) Department of Mechanical and Materials
Engineering
Queen’s University
Kingston, Ontario, CAN

INTRODUCTION

CT-based patient-specific finite element analysis (FEA) is a proposed clinical method to identify osteoporotic patients who could benefit from pharmacological intervention[1]. The largest barrier to implementing this method in the clinic is segmentation. Segmentation is used to isolate phantom plugs in CT scans and create digital geometries. Manual segmentation (MS), the current clinical standard, is time-consuming and only repeatable for skilled operators carefully following a detailed protocol. Thus, there exists a need to improve efficiency and repeatability of segmentation. Alternatively, automated segmentation can be carried out using machine learning (ML) or deep learning (DL) techniques. The aim of this study was to compare the results of ML, DL, and MS in terms contour mesh geometry and density calibration slope.

METHODS

A $Ca_5(PO_4)_5$ calibration phantom (CIRS, Inc, Norfolk, VA, USA) including custom plug densities [mg/cc] of 100 (part: RDH357Y-23) and 400 (part: RDH362Y-24) and stock plugs of 1000 (part: 06217) and 1750 (part: 06221) were CT scanned in air. Data were captured on GE CT scanner models Discovery CT 750HD and Optima 660 (GE Healthcare, Waukesha, WI, USA) with a standard reconstruction kernel. Scan parameters were consistent with established clinical protocols[2] (slice spacing [mm]: 0.325, 0.625; slice thickness [mm]: 0.625, 1.25; voltage [kVp]: 120; current [mA]:105, 120). Phantom plugs were chosen to be the subjects of this study due to their simple geometry. An overview of the study elements and segmentation methods are detailed in Figure 1.

In Dragonfly v. 2020.1 (Object Research Systems Inc., Montreal CAN), multiple ML and DL algorithms were trained to segment plugs from a training dataset of 15 to 20 segmented frames.

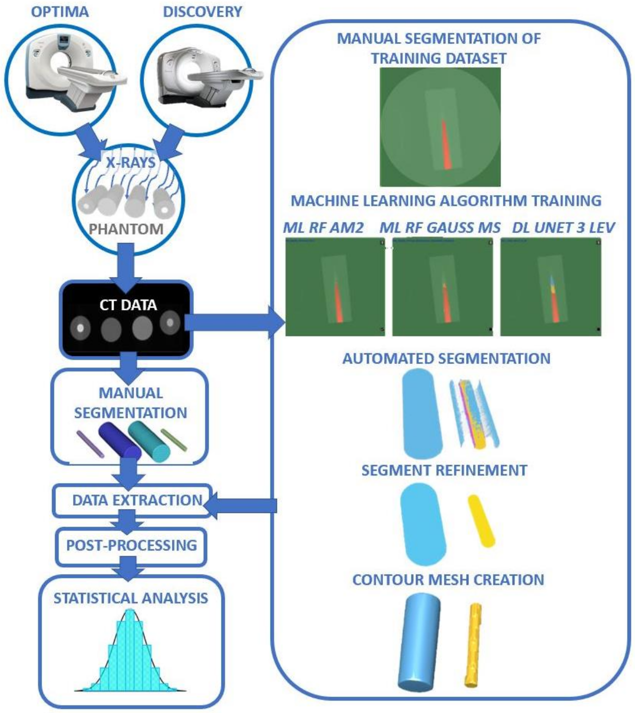

Figure 1: An overview schematic of study elements and parallel segmentation methods. Manual segmentation (MS) (Left), machine learning (ML) and deep learning (DL) segmentation (Right) are detailed.

Algorithm training and related segmentations were performed on a laptop equipped with an NVIDIA GeForce MX 230 GPU. Training was stopped when there were no changes to the loss function monitoring the performance for five iterations. Model performance was assessed based on the percentage of correctly segmented pixels in each training frame and visually by the quality of segments created by each ML segmentation model for an entire CT image stack. The most accurate ML algorithm, *Random Forest Activation Map 2 (ML: RAM)*, was applied to the CT data, to automatically segment a CT image stack of four plugs. Individual plug segments were extracted and artifacts outside of the plug were removed manually. A contour surface mesh of each plug was created from the refined meshes.

3-D geometry and volumetric statistics including standard deviation of CT Number and mean CT Number, were extracted from contour surface meshes for statistical analysis. For comparison purposes, the same types of data were extracted from MS in both Mimics v. 23 (Materialise, Leuven, BE) and Slicer v. 4.10.2 (Slicer, Boston, MA, USA). Geometrical differences in contour surface meshes were compared for results produced by ML, DL and MS using quantiative difference maps in Dragonfly. Density calibration slopes which linearly relate extracted mean CT Numbers [HU] to plug densities [mg/cc] were tested for significant differences between ML, DL, and MS using Mann-Whitney (α= 0.01).

RESULTS

ML segmentation outperformed DL and produced similar results to MS in commercial software. Both ML algorithms had higher pixel accuracy than DL with all algorithms having pixel accuracies greater than 95% (Figure 2). Despite having high pixel accuracies, ML: RGM and DL: U3 failed to correctly segment entire CT image stacks (Figure 3). The ML: RAM and MS in Mimics methods did not produce significantly different slopes (p = 0.097). The ML: RAM and MS in Slicer methods did produce significantly different slopes (p = 0.001).

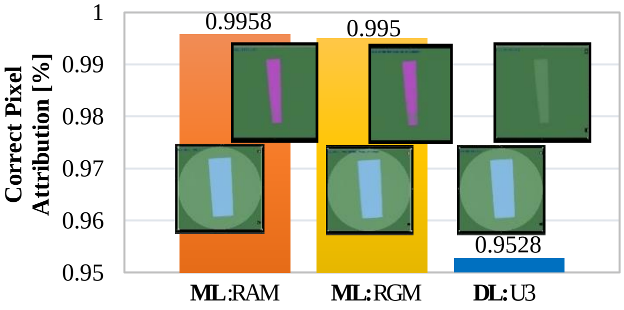

Figure 2: Accuracy measured as correct pixels are plotted for each algorithm. Overlaid are segments of two different plugs produced by machine learning (ML) and deep learning (DL); Random Forest Activation Map 2 (RAM), Random Forest Gaussian-MS (RGM), and 3-level U-Net (U3). These plugs show the varying accuracy of segmentation for each algorithm.

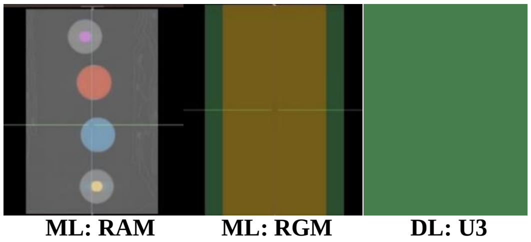

Figure 3: Only ML: RAM accurately performed segmentation on the phantom plugs. A trained operator can discern that ML: RGM and DL: U3 failed to correctly segment the plugs despite high pixel accuracy.

Surface contour meshes created by ML and by MS in Slicer were highly similar for low density plugs (<5% different) as shown in Table 1. Additionally, contour meshes created by ML and by MS in Mimics were very similar (<5% different) for plugs with densities 400 mg/cc or greater. When ML was compared to MS in Slicer and MS in Mimics, standard deviations of geometrical difference were smaller for high density plugs.

Table 1: Plug densities [mg/cc], mean difference, the distance between points [mm], and standard deviation are tabulated. Percent difference is difference normalized by plug diameter.

	Mimics - Dragonfly				Slicer - Dragonfly			
Plug Dens. [mg/cc]	100	400	1000	1750	100	400	1000	1750
Mesh Mean Dif. [mm] *(%)*	7.6 *(19)*	-1.7 *(-4.2)*	-0.3 *(-2.6)*	-0.2 *(-1.5)*	0.5 *(1.2)*	-1.69 *(-4.2)*	-1.7 *(-17)*	-1.8 *(-18)*
Dif. St. Dev. [mm]	10	12	6.8	3.1	13	0.7	0.6	0.8

It took 40 minutes to run ML: RAM for 5mm slice thickness, two hours for 1.25 mm slice thickness, and over four hours at 0.625 mm slice thickness. After which, ten minutes of operator input was necessary to finalize results. Comparatively, trained operators using established protocols for MS in Mimics and MS in Slicer took ten and twelve minutes respectively to complete segmentation.

DISCUSSION

The aim of this study was to compare the results of ML, DL, and MS in terms contour mesh geometry and density calibration slope. ML: RAM performs accurate segmentation with little user input, only requiring ten minutes of trained operator time compared to twelve minutes for an experienced operator to perform MS. However, ML: RAM required up to four hour of run time to segment one image stack of plugs. Higher power workstations intended for research or clinical use will likely perform segmentation tasks more quickly. Additionally, the current time investment for training and running is expected to reduce as technology develops. When compared to ML: RAM, DL: U3 models produced less accurate results, as noted by a trained operator. The poor performance of DL in this setting may be the result of using a small training dataset of 20 or less segmented images. DL is expected to perform better with more training data, which should be tested in future work. Future works should continue to examine the potential error introduced by ML segmentation. A potential follow-up study would be to calculate and compare the impact of ML: RAM and MS in a clinical context such as CT-based patient-specific finite element analysis.

ACKNOWLEDGEMENTS

The authors acknowledge Emily Cameron for prior segmentation used for comparison. We acknowledge the support of the UW-Madison Sophomore Research Fund, Natural Sciences and Engineering Research Council of Canada (NSERC), Ploeg’s Research Initiation, and the Human Mobility Research Centre, Queen’s University, Kingston, ON, Canada. Thanks to Dr. P.J. Pickhardt and Carrie Bartels for help capturing the image data used.

REFERENCES

[1] Imai K. *World J Exp Med*. 5(3):182-187. (2015)
[2] Lee SJ, *Am J Roentgenol*.;209(2):395-402. (2017)