SIIM CMIMI24 Oct 21-22 | Boston, MA
Conference On Machine Intelligence In Medical Imaging

Impact of Random Prostate Volume Modifications on Automated Segmentation Model Performance

Dominic LaBella, MD, Duke University Medical Center; Michaela Kop; Xuan Qi, PhD; Thomas Sanford, MD

Introduction/Background

Accurate and consistent prostate segmentation on magnetic resonance imaging (MRI) plays an important role in surgical planning, biopsy targeting, and radiotherapy planning. Interobserver variability is an inherent challenge for defining a consistent ground truth (GT) segmentation. This study investigates the effect of modifying training set GT segmentations on the performance of automated prostate segmentation models.

Methods/Intervention

Ground truth segmentations of the whole prostate from T2-weighted MRI prostate sequences were manually delineated by a board-certified urologist who is also fellowship trained in urologic oncology. GT segmentation modifications were made by either adding or subtracting a specified distance in millimeters radially from the surface of the whole prostate volume on each axial slice. Each axial slice had a one-third chance of either A) subtracting a uniform inner margin from the prostate surface, B) not modifying the slice's segmentation, or C) adding a uniform outer margin from the prostate surface; as shown in Fig. 1. Ten different modified prostate segmentation-image pair training datasets were created. Each training dataset had a specified amplitude of potential margin modification. Identical segresnet models from the Auto3DSeg framework were trained over 300 epochs for each of the 10 modified training datasets and an additional unmodified GT training dataset. Validation and testing sets included unmodified GT segmentations. Dice similarity coefficients (DSC) were used to compare model performance.

Results/Outcome

A total of 119 T2-weighted images with whole prostate segmentations were included in the study. A linear decrease in mean test set DSC ranged from 0.917 to 0.856 as GT variability increased from 0 to 10 millimeters, as shown in Fig. 2 and Table 1, ($p = 0.0001$, Pearson Correlation Coefficient).

Conclusion

The decrease in testing set DSC as training set segmentation modification amplitude increases shows the importance of consistent GT segmentations for automated segmentation model development. Future studies should assess the impact of interobserver variability across additional radiographical structures on larger datasets.

Statement of Impact

This study elucidates the critical role of consistent GT segmentations when training automated segmentation models. By demonstrating that even minor modifications to GT segmentations can degrade the performance of automated segmentation models, we highlight the importance of standardized segmentation protocols.

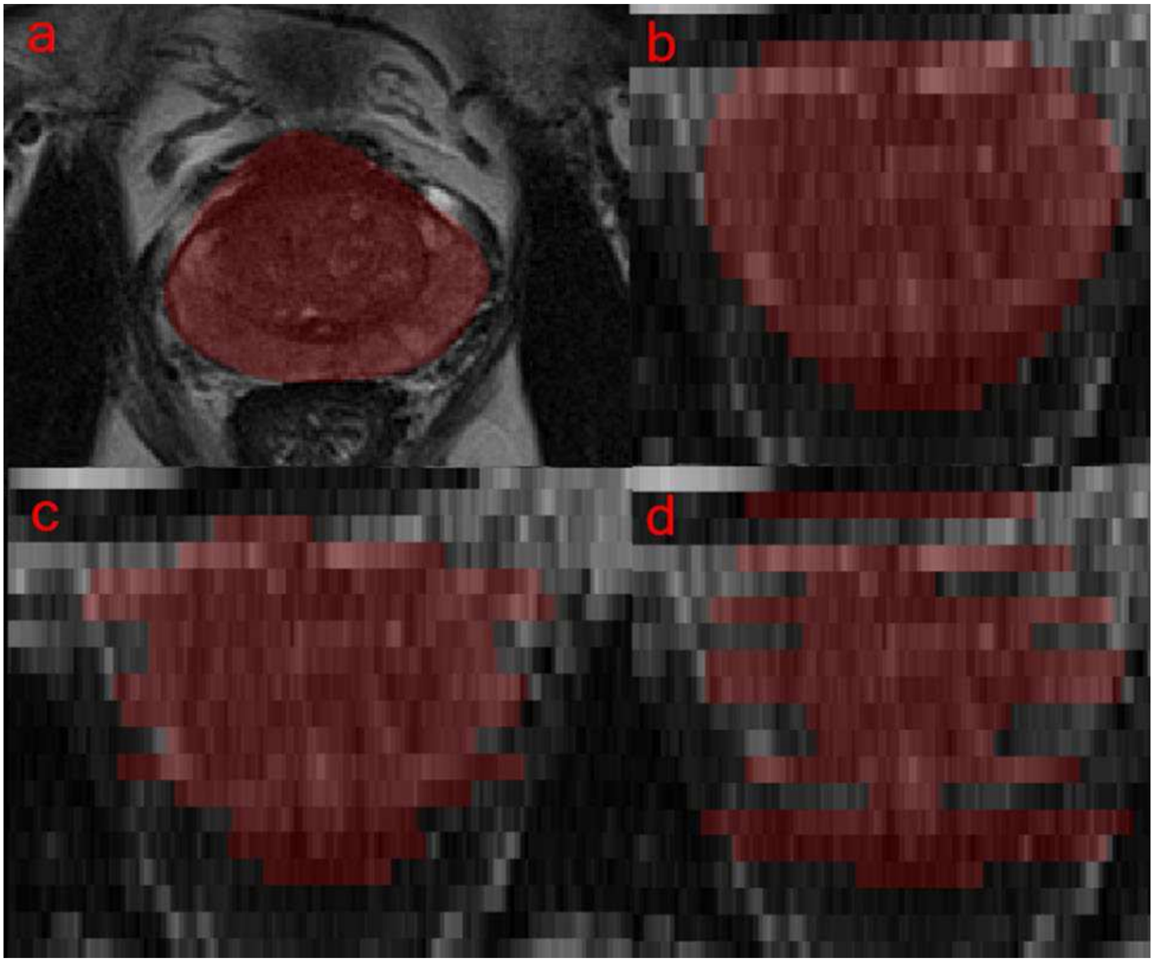
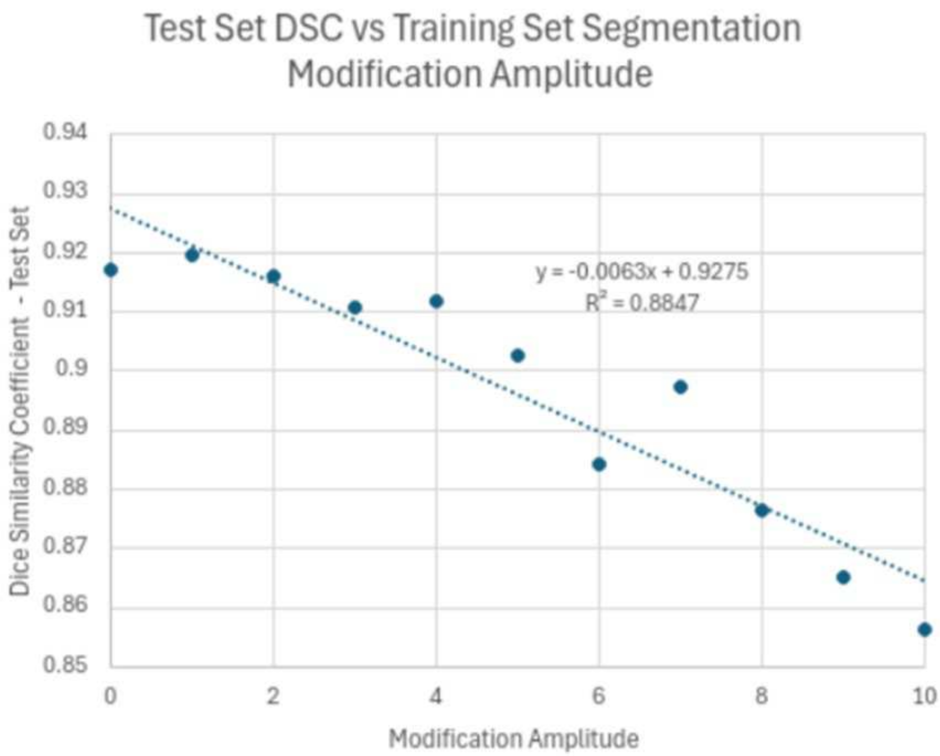


Image panels a-d denote prostate volume segmentations; (a) unmodified ground truth segmentation on a T2-weighted axial image; (b) unmodified ground truth segmentation on a T2-weighted coronal image; (c) training set segmentation with a potential 5 millimeter modification in either the inner or outer radial direction on each axial slice as seen on a T2-weighted coronal image; (d) training set segmentation with a potential 10 millimeter modification in either the inner or outer radial direction on each axial slice as seen on a T2-weighted coronal image.



Scatterplot showing the relationship between the dice similarity coefficient of the automated segmentation models' testing set predictions with the amplitude of the respective training sets' inner and outer radial margin axial-based random modifications.

Modification Amplitude (mm)	Validation Set Average DSC	Best Epoch	Test Set Average DSC
0	0.921	257	0.917
1	0.9188	252	0.9194
2	0.915	290	0.9157
3	0.9159	234	0.9106
4	0.9095	246	0.9115
5	0.907	240	0.9025
6	0.8971	246	0.8839
7	0.8927	152	0.8971
8	0.8813	240	0.8764
9	0.8574	168	0.8649
10	0.8626	160	0.8561

This table presents the average dice similarity coefficient of the validation and test sets, as well as the best epoch for each model, for each respective modification amplitude training set. The datasets were generated with varying amplitudes of ground truth segmentation modifications, ranging from 0 to 10 millimeters.

Keywords
artificial intelligence; automated segmentation; interobserver variability; prostate; machine learning; segresnet