Print this Page for Your Records Close Window

Control/Tracking Number: 2024-SP-4121-RSNA
Activity: Scientific Presentations
Current Date/Time: 5/1/2024 9:20:36 AM

ARE WE ON THE RIGHT WAY FOR EVALUATING AI ALGORITHMS FOR MEDICAL SEGMENTATION?

Author Block: P. Bassi[1], W. Li[2], Y. Tang[3], S. Decherchi[4], A. Cavalli[5], A. Yuille[6], Z. Zhou[2];
[1]Genova, ITALY, [2]Baltimore, MD, [3]Bellevue, WA, [4]Genoa, ITALY, [5]St-Sulpice (VD), SWITZERLAND, [6]Johns Hopkins University, Baltimore, MD.

Abstract:
***Purpose:** To provide a large-scale, open benchmark across 44 advanced AI algorithms in anatomical structure segmentation across three critical dimensions, including accuracy, efficiency, and robustness.
***Methods and Materials:** We have established a large-scale AI benchmark attracting 44 international research teams from 12 countries. These teams develop their AI algorithms, which they either authored first or last, using our expansive annotated CT database. In total, the database consists of 22,682 annotated CT volumes (i.e., 7.9M images) collected from 94 hospitals and 19 countries, where 5,195 out of 22,682 CT volumes are available for training AI algorithms, the remaining are reserved for external validation. Our benchmark evaluates a wide range of AI algorithms, categorized as CNN, Transformer, Mamba, & Vision-Language. In addition, we also evaluate publicly available AI frameworks—which are more flexible and can support different backbones—in terms of their segmentation capacity. These frameworks include Auto3DSeg from NVIDIA, nnU-Net from DKFZ, and numerous other open-source repositories developed by researchers. These models and frameworks, which collectively have garnered over 102,000 citations, are evaluated independently by our team using an unseen dataset of 13,420 CT volumes from 12 hospitals to determine their performance measured by segmentation accuracy (DSC and NSD) and inference efficiency (running time per case). Our benchmarking effort is featured in challenges hosted by ISBI-2024 and MICCAI-2024, invited as a Lighthouse Challenge in MICCAI-2025, and is planned to expand over the next five years.
***Results:** We have received 15 AI algorithms and expect at least 35 more. DSC scores vary from 82% (SegVol) to 92% (UniSeg) across AI algorithms, 70% (aorta) to 96% (liver & spleen) across anatomical structures, and inference times range from 0.12 (SegVol) to 0.61 (Diff-UNet V2) seconds per CT slice.
***Conclusions:** Our AI benchmark addresses critical limitations in existing medical AI evaluations by ensuring that: (I) creators of 44 renowned AI algorithms use consistent training and testing splits to mitigate biased reproductions by other teams, (II) our team conducts all evaluations independently, prohibiting test-time adjustments that could lead to overfitting, and (III) the test set comprises a comprehensive, diverse, and entirely unseen collection of 13,420 CT volumes from 12 hospitals, making it the largest test set used to date.
***Clinical Relevance/Application:** The outcome will be translated into clinical prototypes. These AI prototypes have been deployed and beta-tested in JHU and UCSF to assist radiologists in automated organ volume measurement and support surgeons in precise surgical navigation.

Category (Complete): Imaging Informatics -> INRR - Results and Reporting
Format Preference (Complete): Oral Paper
Questions (Complete):
Trainee Research Prize: Not Applicable
Disclosure of "Off-Label" usage: No, I do not intend to discuss off-label uses
IRB / IACUC Response: Human subject, and received IRB approval
Has this work been previously presented or published?: No
2nd Format Opportunity: Yes, I would be interested if accepted to showcase my work in an additional format (2 meter Hardcopy Backboard).

Attached Files: We established a large-scale AI benchmark attracting more than 50 global research teams, driving innovation from classical methods (U-Net and its variants) to cutting-edge Foundation Models. (PDF, 930489 bytes)

Status: Complete

Feedback