SIIM CAIMI25
Conference on Artificial Intelligence in Medical Imaging
OCT 20-21 | UNIVERSITY OF CALIFORNIA, IRVINE

# Turning Noise into Signal: Interpretable Disagreement Profiling for Improved AI-Based 3D Tumor Segmentation

Fabian Umeh, MSc, Teesside University; Monika Pytlarz, MSc; MingDe Lin, PhD; Nazanin Maleki, MD; Raisa Amiruddin, MBBS; David Weiss; Khaled Bousabarah; Marina Ivory, MD, PhD; Mohamed Ghonim, MD; Mohanad Ghonim, MD; Albara Alotaibi, MD; Melisa Guelen, MD, PhD; Nathan Page, MD; Pascal Fehringer; Sedra Mhana; Bojan Petrovic, MD; Fatima Memon, MD, MPH; Basimah Albalooshy, MD; Elizabeth Schrickel, MD; Justin Cramer, MD; Michael Veronesi, MD, PhD; Spyridon Bakas, PhD; Mariam Aboian, MD, PhD

**Introduction/Background**
As oncology treatment response guidelines increasingly leverage technology for more comprehensive tumor characterization and longitudinal tracking, the need for accurate and workflow-efficient 3D tumor segmentation becomes critical. AI segmentation models are limited by inter-rater variability in their manual training annotations, an issue that can limit model performance. In this work, we introduce a novel framework that reframes inter-reader variability not as noise to be minimized, but as a valuable signal. This is achieved through two key innovations: first, the creation of a detailed “Disagreement Profile” that captures the nature of annotation variability; and second, the use of this profile to actively guide the generation of a more robust consensus segmentation from multiple raters.

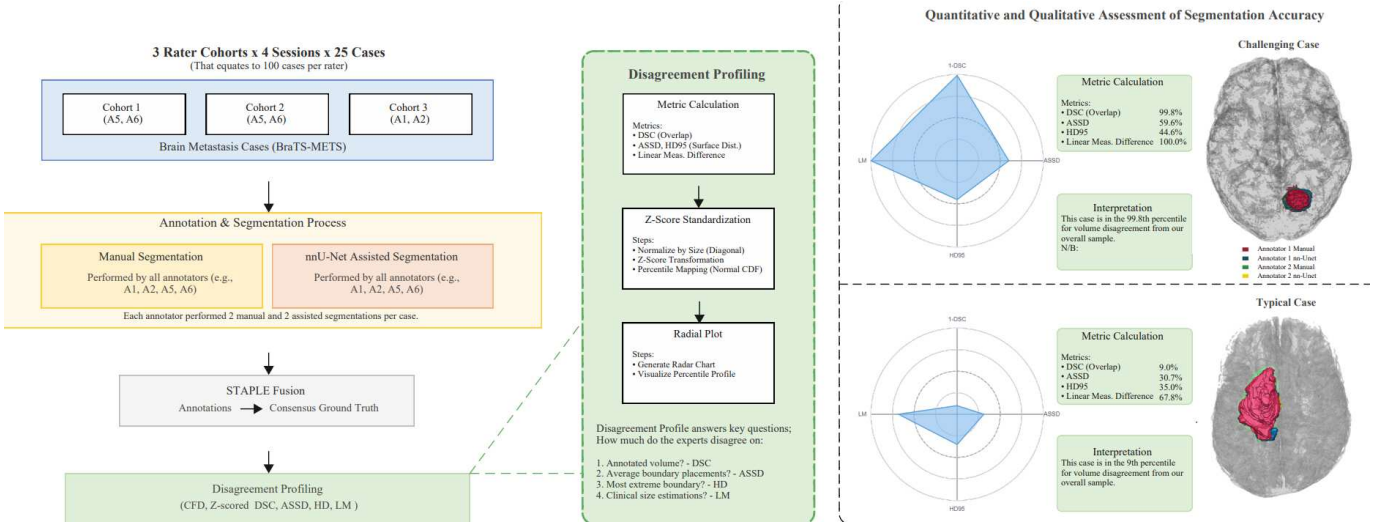
**Methods/Intervention**
We evaluated a cohort of brain metastases patients, each segmented eight times by six raters. In our two-stage methodology, we first generated Disagreement Profiles by comparing each rater's annotation to a STAPLE consensus. For comparability, our four metrics (DSC, ASSD, HD95, LM) were converted to percentile ranks using cohort-specific z-scores and the normal cumulative distribution function (CDF), rather than an idealized standard. In the second stage, we developed an iterative, weighted-mean consensus-building algorithm guided by the Disagreement Profiles. This algorithm selectively amplifies the metric-specific strengths of individual annotators to generate a final, robust consensus.

**Results/Outcome**
Using a leave-one-out cross-validation approach across three cohorts (n=75), the profile-guided consensus method demonstrated a statistically significant improvement in geometric segmentation metrics compared to a standard STAPLE baseline. Our method produced a significantly higher mean DSC (0.90 ± 0.09 vs. 0.88 ± 0.11, $p < 0.001$) and a significantly lower ASSD, indicating better boundary agreement (1.76 ± 3.14 mm vs. 2.50 ± 4.95 mm, $p = 0.008$). Z-scores effectively identified outlier cases and generated interpretable disagreement “signatures” for each. Visualizing these signatures as radial plots offered an intuitive view of multi-dimensional rater divergence.

**Conclusion**
Our two-stage framework powerfully analyzes multi-rater variability in neuro-oncology imaging. By converting disagreement into an actionable profile that guides the consensus process, it produces more reliable and confident reference segmentations.

## Statement of Impact

Our profile-guided approach enhances AI training data quality and fosters more robust, reproducible standards for imaging-based clinical research.




Disagreement Profile: This profile is designed to quantify the variability between different raters' segmentations. It is composed of four distinct metrics, each assessing a different aspect of disagreement: • Volumetric Overlap: Measured by the Dice Similarity Coefficient (DSC). • Average Boundary Error: Measured by the Average Symmetric Surface Distance (ASSD). • Worst-Case Boundary Discrepancies: Measured by the 95th-percentile Hausdorff Distance. • Clinical Size Estimation: Measured by a Linear Measurement. When this profile is displayed on a radial plot, the dashed line serves as a visual baseline, representing the average or median performance within the dataset (50th percentile, equivalent to a Z-score of 0) for each metric.

**Algorithm 1 Consensus and Z-Score Computation**

**Require:** Case with N rater masks $M = \{m_1, \ldots, m_n\}$
**Ensure:** Z-scores $z_C$ for case C

Compute STAPLE consensus map P; binarize at 0.5 to get B.
**if** no consensus ($\Sigma B = 0$) **then return** {NaN}[4]
Calculate metrics (DSC, ASSD, HD95, LM_diff) for each mask vs. consensus B.
**for** each metric k **do**
    Compute mean $\mu_k$ across raters.
    Score = (k = DSC) ? 1 - $\mu_k$ : $\mu_k$ / BoundingBoxDiagonal(B).
**end for**
Standardize scores to Z-scores across all cases.

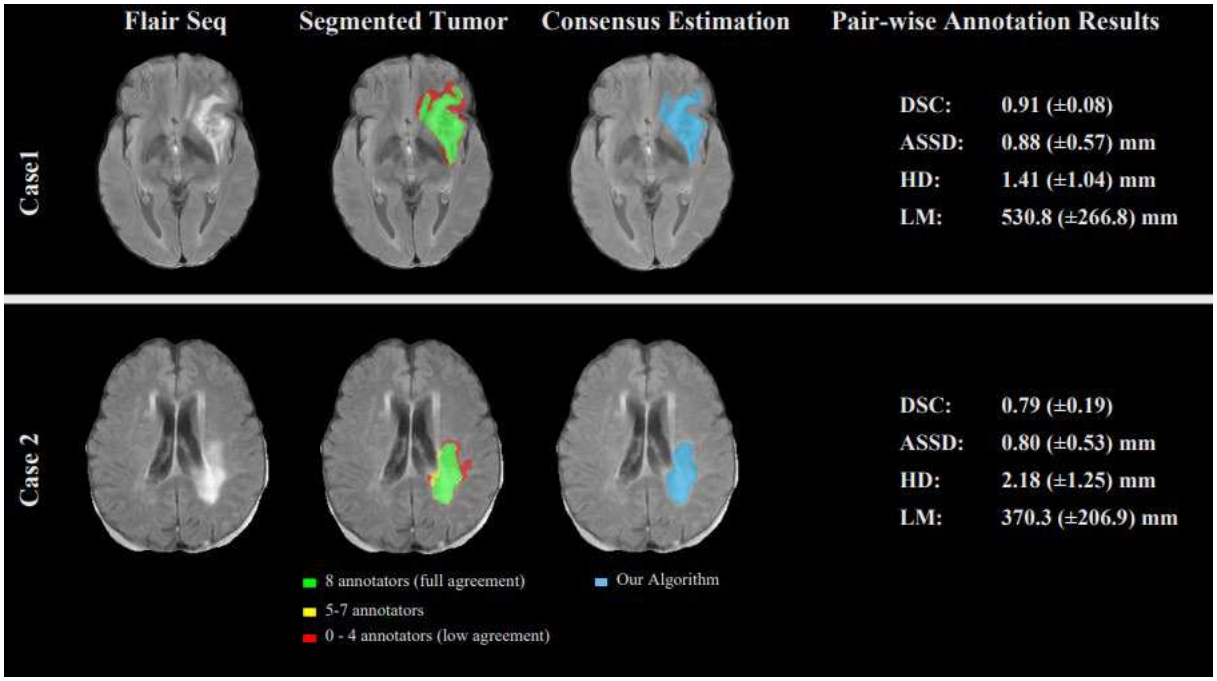
**Algorithm 2 Iterative Weighted Mean**

**Require:** Set of N masks $M = \{m_1, ..., m_N\}$
**Require:** Set of N Z-score vectors $Z = \{z_1, ..., z_N\}$ (optional)
**Ensure:** Final consensus map $P_{final}$

$P_{ref} \leftarrow I(\ (1/N)\ \Sigma_{i=1 \text{ to } N}\ m_i \geq 0.5\ )$ // Majority vote
**for** t=1 to $T_{max}$ **do**
    $B_{ref} \leftarrow I(P_{ref} > 0.5)$
    **for** i=1 to N **do**
        $s_i \leftarrow$ NormalizeMetrics($m_i$, $B_{ref}$)
        **if** Z is provided **then**
            $w_{metric,i} \leftarrow \sigma(-z_i)$
        **else**
            $w_{metric,i} \leftarrow [1/4, 1/4, 1/4, 1/4]$
        **end if**
        $c_i \leftarrow s_i \cdot w_{metric,i}$ // Dot product
    **end for**
    $w_{ann} \leftarrow \sigma(c)$ // Softmax over composite scores vector
    $P_{new} \leftarrow \Sigma_{i=1 \text{ to } N}\ w_{ann,\,i} \cdot m_i$
    **if** $\max(|P_{new} - P_{ref}|) < \varepsilon$ **then**
        **break**
    **end if**
    $P_{ref} \leftarrow P_{new}$
**end for**
**return** $P_{ref}$

Algorithm 1: Disagreement Profile Generation This algorithm generates a Disagreement Profile to quantify inter-rater variability for each case. First, a baseline consensus is derived using the STAPLE algorithm. Each individual annotation is then compared against this consensus across four metrics: volumetric overlap (Dice Similarity Coefficient), average boundary error (ASSD), worst-case boundary discrepancies (95th-percentile Hausdorff Distance), and clinical size estimation (Linear Measurement). Finally, these raw metrics are converted into cohort-specific z-scores to create a standardized profile for robust, relative comparison. Algorithm 2: Guided Consensus Generation This SoftMax-weighted algorithm dynamically adjusts the influence of

each annotation based on its z-scores across four metrics. Annotations with more favourable (i.e., negative) z-scores for a specific metric contribute more heavily to the final consensus in that dimension. For instance, an annotation with minimal boundary error (a low ASSD z-score) becomes more influential in shaping the consensus boundary, effectively leveraging the specific strengths of each rater.




Flair Seq: A representative axial FLAIR sequence image of patients with brain metastasis. Segmented Tumour: A visual map of annotation agreement among eight independent raters. Green indicates high agreement (i.e., a high number of raters included the voxel), while red indicates low agreement at the tumour boundary. Consensus Estimation: The final, single segmentation mask generated by our proposed profile-guided algorithm. The pair-wise annotation metrics, such as mean and standard deviation, were calculated for each case by comparing all 28 possible pairwise combinations of the eight annotators' masks.


## Keywords

Consensus Segmentation; Inter-Rater Variability; Brain Metastases; Medical Image Analysis; Disagreement Profile; Artificial Intelligence (AI)