**Supplemental File 16**: Detailed Qualitative Assessments of Deep Learning Segmentation Masks by Thoracic Radiologists

This supplemental file presents the comprehensive raw data from the qualitative visual assessments conducted by six board-certified thoracic radiologists (D1-D6), each with 5–15 years of clinical experience. The evaluations focus on segmentation masks produced by five deep learning models—3D Attention U-Net, ResUNet, VNetVNet, ReconNet, and SAM-Med3D—applied to lung cancer CT images. Responses are compiled across 21 questions (q1–q21), categorized into seven domains: overall impression, spatial differences, tumor heterogeneity and boundary quality, survival prediction relevance, trust and preference, summaries and overall assessment, and workflow and adoption. For questions q1–q13 and q15–q20, individual Likert-scale ratings (typically 1–5, with higher scores indicating more favorable outcomes except for q5 where lower is better) are provided, along with any accompanying descriptive comments. Question q14 captures open-ended insights on requirements for trustworthy AI segmentation, while q21 records categorical preferences for AI's role in clinical practice (Options A: Replace physicians entirely; B: Provide an initial mask for refinement; C: Assist as a support/check tool). Data are organized in tabular format by radiologist, model, and question to facilitate transparency and reproducibility, supplementing the summarized findings in the main manuscript's Qualitative Visual Assessment section.

**D1**

# **Comparative Clinical Assessment of Deep Learning Tumor Masks vs. Ground Truth in Lung Cancer**

Model under review: 3DAttention

**Definitions**

**Prediction Evaluation Metric Definition**

* **Accuracy.** The proportion of all predictions (both positive and negative) that are correct out of the total cases.
* **Precision:** Of all the cases predicted positive, how many were actually positive.
* **Recall (Sensitivity):** Of all the actual positive cases, how many were correctly identified.
* **F1 Score:** The balance between precision and recall, combining them into one number.
* **AUC (Area Under the Curve):** Measures overall ability of the model to separate classes, regardless of threshold.

**Segmentation Evaluation Metric Definition**

* **Dice Score (Dice Similarity Coefficient):** Measures overlap between two regions (e.g., predicted vs. true mask); higher values mean better overlap.
* **IoU Score (Intersection over Union / Jaccard Index):** Ratio of the common area (intersection) to the total area (union) of prediction and ground truth.
* **Hausdorff Distance:** Measures the largest boundary mismatch between two shapes; lower values mean closer alignment.

**Segmentation Results**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Analysis Mode | Model | Dice Score | IoU Score | Hausdorff Distance |
| Test |  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
| **3DAttention** | 0.77 0.08 | 0.64 0.1 | 33.61 4.10 |

**Prediction Result:**

Radiomics extracted from Ground Truth used for prediction tasks

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Validation | | | | | | Test | | | | | | | Names | | | |
| Acc | Precision | Recall | F1 | AUC | Specificity | Acc | Precision | Recall | F1 | AUC | Specificity | FSA/AEA | | CA | Model | SL/SSL |
| 0.88 | 0.78 | 0.88 | 0.83 | 0.54 | 1 | 0.87 | 0.76 | 0.87 | 0.81 | 0.71 | 1 | Mutual\_Info\_Gain\_Ratio | | ExtraTree | GT | SSL |
| 0.74 | 0.67 | 0.74 | 0.66 | 0.55 | 0.98 | 0.87 | 0.79 | 0.87 | 0.83 | 0.63 | 0.99 | SL |

Radiomics extracted from DL 3D attention used for prediction tasks

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Validation | | | | | | Test | | | | | | | Names | | | |
| Acc | Precision | Recall | F1 | AUC | Specificity | Acc | Precision | Recall | F1 | AUC | Specificity | FSA/AEA | | CA | Model | SL/SSL |
| 0.79 | 0.80 | 0.79 | 0.80 | 0.69 | 0.87 | 0.81 | 0.83 | 0.81 | 0.82 | 0.76 | 0.88 | Guassian Random Projection | | Voting Classifier | 3DAtt | SSL |
| 0.61 | 0.64 | 0.61 | 0.62 | 0.53 | 0.69 | 0.72 | 0.79 | 0.72 | 0.75 | 0.55 | 0.79 | SL |

**Start**

## Section 1 — Overall Impression

1.1 Does the AI mask define a tumor volume that is more clinically meaningful than the Ground Truth (GT) for diagnosis and visual interpretation?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ■ 5 ☐ (0 = not at all, 5 = very strongly yes)

Reason: The AI mask reliably captures not only the solid tumor core but also the surrounding ground-glass opacity (GGO) and peritumoral regions. Although this larger volume is less aligned with the standard Gross Tumor Volume (GT), it seems to include more biologically relevant information, which is reflected in its better performance in the downstream survival prediction task (AUC 0.76 vs. 0.71).

## Section 2 — Spatial Differences

2.1 Are the AI–GT differences concentrated in the peritumoral zone (≈5–15 mm around the tumor)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ■ (0 = not at all, 5 = strongly concentrated)

Reason: Yes, the visual review of all cases shows a clear and consistent pattern where the main discrepancy is the AI mask's inclusion of a margin around the GT. This over-segmentation is mainly localized to the peritumoral zone and relates to areas of ground-glass opacity (GGO) and possible microscopic invasion, rather than random spatial errors.

2.2 How clinically meaningful are the regions where the AI mask differs from GT (including over-segmentation or under-segmentation, such as peritumoral expansion, missed necrotic cores, spiculated margins)? Or, how clinically justified are the AI–GT differences (whether due to additional regions or excluded regions)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ■

Reason: The differences between AI-GT, which show consistent over-segmentation in the peritumoral zone, are clinically significant. These areas, corresponding to ground-glass opacities (GGOs), likely represent the tumor's microenvironment, including microscopic invasion and inflammation. Features from the AI's larger volume yield better prognostic results (AUC 0.76) than the more conservative GT mask (AUC 0.71). The AI's 'error' is actually a justified expansion to a more complete biological target volume.

## Section 3 — Clinical Interpretation of Differences

3.1 How well does the AI mask preserve internal tumor heterogeneity (for example: solid vs. ground-glass opacity, necrotic vs. viable regions)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ■

Reason: The AI mask effectively maintains heterogeneity by capturing both the solid tumor core and the surrounding peritumoral GGO. Defining a volume that encompasses various tissue textures provides a more comprehensive view of the lesion's internal complexity compared to the GT, which primarily focuses on the solid component.

3.2 Boundary quality: compared to GT, how noisy or fragmented are AI boundaries?

0 ☐ 1 ■ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = much less noisy, 5 = much more noisy)

Reason: The AI-generated boundaries are not noisy or fragmented. They are consistently smooth and coherent contours. Although the mask is larger than the GT, this is a systematic expansion, not a result of a noisy or unstable boundary-detection process.

## Section 4 — Link to Survival Prediction

4.1 Do the AI–GT differences likely capture prognostically relevant tissue that improves survival prediction?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ■

Reason: The quantitative results demonstrate improved survival prediction with the AI mask (AUC 0.76 vs. 0.71). Visual analysis confirms that the AI-GT difference is due to the consistent inclusion of the peritumoral zone, indicating that AI captures prognostic tissue that is missed by standard GT delineation.

4.2 How would you rate the effect of the AI’s deviation from GT (larger or smaller volume) on survival prediction?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ■ (0 = strongly harmful, 5 = strongly beneficial)

Reason: The AI's deviation (a larger volume including the peritumoral halo) is highly valuable for survival prediction. By expanding beyond just the solid tumor mass (GTV) to a broader biological target volume, the AI provides a richer set of features for the prognostic model, leading to a significantly more accurate prediction of patient outcomes.

## Section 5 — Trust & Preference (Survival Prediction Focus)

5.1 How likely is it that AI-generated masks are more accurate than GT in some cases (for example, subtle lesions missed by physicians, or over-delineation that reduces noise) in survival prediction tasks?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ■ (0 = very unlikely, 5 = very likely)

Reason: Since the AI mask-derived model has already demonstrated better prognostic performance (AUC 0.76 vs. 0.71), its delineation is likely more accurate for survival prediction. This is not a random occurrence but a systematic feature, as the AI consistently captures the prognostically critical peritumoral zone, which is often excluded from standard GT segmentations.

5.2 Would you prefer to use the AI mask (rather than GT) as input for survival prediction?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ■ (0 = not at all, 5 = strongly yes)

Reason: For the specific purpose of developing a survival prediction model, the AI mask is the preferred input. The higher AUC score demonstrates its superiority. It offers a more comprehensive "biological tumor volume" that better reflects the tumor's interaction with its microenvironment, resulting in a more accurate and clinically valuable prognostic assessment.

## Section 6 — Summaries and Overall Assessment

6.1 Diagnosis-oriented evaluation — How successful is the AI mask in lesion delineation compared to GT?

0 ☐ 1 ☐ 2 ☐ 3 ■ 4 ☐ 5 ☐ (0 = complete failure, 5 = highly successful)

Reason: For the specific task of outlining the solid Gross Tumor Volume (GTV), the AI mask is only moderately effective. While it correctly identifies the tumor's core, its consistent over-segmentation into the peritumoral zone makes it less precise as a geometric match to the GT compared to other models.

6.2 Prediction-oriented evaluation — How successful is the AI mask in contributing to survival prediction compared to GT?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ■ (0 = no added value, 5 = very strong added value)

Reason: The AI mask provides significant value for survival prediction, outperforming the GT (AUC 0.76 vs. 0.71). Its success comes from capturing a broader 'biological tumor volume' that includes the prognostically important peritumoral microenvironment, which is crucial for accurate outcome modeling.

6.3 Overall confidence in your assessment

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ■ (0 = no confidence, 5 = complete confidence)

Reason: Confidence is complete because the conclusions are supported by both quantitative data (superior AUC for the AI mask) and consistent qualitative findings across all reviewed cases, which provide a clear clinical explanation for the quantitative results.

6.4 Do you think geometric evaluation metrics (e.g., Dice, IoU, Hausdorff distance) are sufficient to establish clinical trust in AI segmentation?

0 ■ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not sufficient at all, 5 = fully sufficient)

Reason: No, they are not sufficient. This model, which had the worst Hausdorff distance, produced the best survival prediction. This shows that geometric metrics alone can be misleading and fail to reflect the clinical or biological significance of an AI segmentation, which may be "inaccurate" but more useful.

6.5 What additional types of evaluation do you think are necessary for trustworthy AI segmentation (e.g., radiomics stability, survival prediction, physician review, workflow impact, and others), and why do you think those are effective in building trustworthiness?

Reason: For AI segmentation to be trustworthy, evaluation must go beyond geometric metrics. Significant additions include: 1) \*\*Downstream Task Performance\*\*, such as its effect on survival prediction or treatment response models, to assess actual clinical usefulness. 2) \*\*Expert Physician Review\*\*, to interpret the clinical significance of AI-GT discrepancies (e.g., whether an 'error' actually detects a meaningful biological feature). 3) \*\*Radiomics Feature Stability Analysis\*\*, to understand how segmentation differences impact the quantitative features used in predictive models.

## Section 7 — Clinical Adoption & Workflow Impact

7.1 Usability — How easy is it to interpret and use the AI mask compared to GT in clinical workflow?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ■ 5 ☐ (0 = much harder, 5 = much easier)

Reason: The AI mask is simple to interpret because it behaves predictably. Although it is different from a standard GTV, its consistent use of the peritumoral halo is an easy concept for clinicians to grasp and include in their assessment.

7.2 Clinical actionability — Would the AI–GT differences change clinical decisions (for example, radiotherapy target, surgery extent, follow-up planning)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ■ (0 = no impact, 5 = strong impact)

Reason: Differences between AI-GT could significantly influence clinical decisions. In radiotherapy, the AI's volume can help determine the expansion from a Gross Tumor Volume (GTV) to a Clinical Target Volume (CTV), ensuring the area of likely microscopic spread is targeted. In surgery, it might indicate the need for a wider margin.

7.3 Robustness and variability — How consistent and reliable are AI masks across different cases and observers?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ■ (0 = very inconsistent, 5 = very consistent)

Reason: Examining the cases shows that the AI mask is highly consistent. It reliably reproduces the same behavior—capturing both the solid core and the peritumoral halo—across different patients, demonstrating a robust and well-generalized model.

7.4 Time-saving potential — How much time could this AI save compared to manual delineation?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ■ (0 = no time saved, 5 = saves most effort)

Reason: The AI provides a complete 3D segmentation mask instantly. Compared to the laborious and time-consuming process of manual slice-by-slice delineation, the time-saving potential is enormous, freeing up physician time for higher-level tasks.

7.5 Trust in clinical use — How confident would you be using AI masks for high-stakes tasks (e.g., radiotherapy planning) without physician refinement?

0 ☐ 1 ■ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no trust, 5 = complete trust)

Reason: For critical tasks like radiotherapy planning, there would be low trust in using this mask without a physician’s refinement. While its larger volume provides prognostic value, defining a treatment target precisely requires expert clinical judgment. The mask acts as an excellent proposal or a "biological guide," but it cannot replace the final approval and editing by a physician.

7.6 Confidence maps/interpretability (if available) — Do uncertainty or confidence maps improve your trust in AI outputs?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no improvement, 5 = very strong improvement)

Score: N/A (Assuming not available, but providing rationale)

Reason: Confidence maps were not provided for this evaluation. However, if they had been available, they would have greatly increased trust (5 out of 5). An uncertainty map showing the peritumoral zone as having lower confidence than the solid core confirms that the AI recognizes the ambiguity of this region, thereby boosting a clinician's trust in the model's "self-awareness."

7.7 Role of AI in clinical practice — In your opinion, AI-based segmentation should:

☐ Replace physicians entirely  
■ Provide an initial mask for physicians to refine  
■ Assist physicians as a support/check tool

Reason: This model is not intended to replace physicians. Its optimal role is as a powerful assistant. It can generate a highly accurate initial draft of a treatment plan that a physician then refines. Additionally, its unique ability to highlight a prognostically significant biological region makes it a valuable "second read" or support tool, guiding the clinician's attention to areas that need careful review.

**End**

**Comparative Clinical Assessment of Deep Learning Tumor Masks vs. Ground Truth in Lung Cancer**

Model under review: 3DMedSam

**Segmentation Results**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Analysis Mode | Model | Dice Score | IoU Score | Hausdorff Distance |
| Test |  |  |  |  |
|  |  |  |  |
|  |  |  |  |
| **3DMedSam** | 0.68 0.19 | 0.57 0.18 | 4.13 1.55 |

**Prediction Result:**

Radiomics extracted from Ground Truth used for prediction tasks

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Validation | | | | | | Test | | | | | | | Names | | | |
| Acc | Precision | Recall | F1 | AUC | Specificity | Acc | Precision | Recall | F1 | AUC | Specificity | FSA/AEA | | CA | Model | SL/SSL |
| 0.88 | 0.78 | 0.88 | 0.83 | 0.54 | 1 | 0.87 | 0.76 | 0.87 | 0.81 | 0.71 | 1 | Mutual\_Info\_Gain\_Ratio | | ExtraTree | GT | SSL |
| 0.74 | 0.67 | 0.74 | 0.66 | 0.55 | 0.98 | 0.87 | 0.79 | 0.87 | 0.83 | 0.63 | 0.99 | SL |

Radiomics extracted from DL 3DmedSam used for prediction tasks

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Validation | | | | | | Test | | | | | | Names | | | |
| Acc | Precision | Recall | F1 | AUC | Specificity | Acc | Precision | Recall | F1 | AUC | Specificity | FSA/AEA | CA | Model | SL/SSL |
| 0.88 | 0.77 | 0.88 | 0.82 | 0.67 | 1 | 0.85 | 0.71 | 0.85 | 0.77 | 0.68 | 1 | ReliefF | GradientBoosting | MedSam | SSL |
| 0.74 | 0.55 | 0.74 | 0.63 | 0.51 | 0.99 | 0.86 | 0.76 | 0.86 | 0.81 | 0.71 | 0.99 | SL |

**Start**

Section 1 — Overall Impression

1.1 Does the AI mask define a tumor volume that is more clinically meaningful than the Ground Truth (GT) for diagnosis and visual interpretation?

0 ☐ 1 ■ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not at all, 5 = very strongly yes)

Reason: The AI mask is not clinically useful because it consistently and severely under-segments the tumor volume. In all reviewed cases, the mask only captures a small, central part of the lesion, leaving out important portions of the tumor mass. This results in an incomplete and misleading representation of the true tumor burden, which is reflected in its poor performance on the downstream survival prediction task (AUC 0.68).

Section 2 — Spatial Differences

2.1 Are the AI–GT differences concentrated in the peritumoral zone (≈5–15 mm around the tumor)?

0 ■ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not at all, 5 = strongly concentrated)

Reason: The AI-GT differences do not appear in the peritumoral zone. Instead, the discrepancy results from significant under-segmentation, where the AI mask misses most of the actual tumor volume. The error involves omitting intra-tumoral tissue, not adding extra-tumoral tissue.

2.2 How clinically meaningful are the regions where the AI mask differs from GT (including over-segmentation or under-segmentation, such as peritumoral expansion, missed necrotic cores, spiculated margins)? Or, how clinically justified are the AI–GT differences (whether due to additional regions or excluded regions)?

0 ■ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

Reason: The regions of difference, which indicate missed tumor tissue due to severe under-segmentation, are clinically important. The AI's failure to include these regions is not justified clinically and results in a significant underestimation of tumor burden. Omitting meaningful cancerous tissue makes the mask unsuitable for clinical or prognostic use.

Section 3 — Clinical Interpretation of Differences

3.1 How well does the AI mask preserve internal tumor heterogeneity (for example: solid vs. ground-glass opacity, necrotic vs. viable regions)?

0 ■ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

Reason: The AI mask performs poorly in this aspect. It consistently under-segments and only captures a small, central part of the lesion, failing to retain the tumor's overall internal heterogeneity. It completely disregards the textural and structural information from the significant parts of the tumor that it misses.

3.2 Boundary quality: compared to GT, how noisy or fragmented are AI boundaries?

0 ☐ 1 ■ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = much less noisy, 5 = much more noisy)

Reason: The boundary of the AI mask, for the small portion it segments, is smooth and continuous, not noisy or fragmented. This aligns with the very low Hausdorff Distance metric. The model's main issue is its incompleteness, not the boundary quality.

Section 4 — Link to Survival Prediction

4.1 Do the AI–GT differences likely capture prognostically relevant tissue that improves survival prediction?

0 ■ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

Reason: No, the opposite is true. The AI-GT difference is the large volume of tumor tissue that the AI misses. By failing to capture the full extent of the lesion, the model discards prognostically critical information, which directly causes its poorer performance in the survival prediction task.

4.2 How would you rate the effect of the AI’s deviation from GT (larger or smaller volume) on survival prediction?

0 ■ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = strongly harmful, 5 = strongly beneficial)

Reason: The AI's deviation, which results in a significantly smaller volume due to under-segmentation, is highly detrimental to survival prediction. This is supported by the model's lower AUC score (0.68) compared to the GT (0.71). Omitting critical tumor data fundamentally weakens the foundation for an accurate prognostic model.

Section 5 — Trust & Preference (Survival Prediction Focus)

5.1 How likely is it that AI-generated masks are more accurate than GT in some cases (for example, subtle lesions missed by physicians, or over-delineation that reduces noise) in survival prediction tasks?

0 ■ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = very unlikely, 5 = very likely)

Reason: It is very unlikely that this AI mask is more accurate. The model's main trait is consistent and severe under-segmentation, meaning it misses large, clinically important parts of the tumor. This is confirmed by its lower prognostic accuracy (AUC 0.68) compared to the more complete Ground Truth mask (AUC 0.71).

5.2 Would you prefer to use the AI mask (rather than GT) as input for survival prediction?

0 ■ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not at all, 5 = strongly yes)

Reason: I prefer not to use this AI mask. The Ground Truth mask offers a more complete and accurate representation of the tumor volume, leading to a more precise and effective survival model. Relying on the AI mask, which is based on incomplete data, could provide a less reliable and potentially misleading prognostic assessment.

Section 6 — Summaries and Overall Assessment

6.1 Diagnosis-oriented evaluation — How successful is the AI mask in lesion delineation compared to GT?

0 ■ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = complete failure, 5 = highly successful)

Reason: The AI mask completely fails at lesion delineation. Its consistent and severe under-segmentation means it does not accurately capture the tumor's true extent, making it diagnostically unreliable and clinically misleading.

6.2 Prediction-oriented evaluation — How successful is the AI mask in contributing to survival prediction compared to GT?

0 ■ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no added value, 5 = very strong added value)

Reason: The AI mask offers no added value and is actually harmful to survival prediction. By removing large portions of tumor tissue, it excludes critical prognostic information, leading to a model with worse predictive accuracy (AUC 0.68) than the Ground Truth (AUC 0.71).

6.3 Overall confidence in your assessment

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ■ (0 = no confidence, 5 = complete confidence)

Reason: Confidence is high because the visual evidence of severe under-segmentation across all cases matches exactly with the poor quantitative results (low Dice score, poor AUC). The model's failure is consistent and clear.

6.4 Do you think geometric evaluation metrics (e.g., Dice, IoU, Hausdorff distance) are sufficient to establish clinical trust in AI segmentation?

0 ■ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not sufficient at all, 5 = fully sufficient)

Reason: No. This model exemplifies their inadequacy. It had the best (lowest) Hausdorff distance, indicating excellent boundary alignment. However, this metric completely failed to reveal the critical clinical failure of severe under-segmentation, making it highly misleading on its own.

6.5 What additional types of evaluation do you think are necessary for trustworthy AI segmentation (e.g., radiomics stability, survival prediction, physician review, workflow impact, and others), and why do you think those are effective in building trustworthiness?

Reason: To build trust, AI segmentation requires a multi-faceted evaluation beyond simple geometric scores. This must include:

1. Downstream task performance: Evaluating the model's effect on a final clinical goal (e.g., survival prediction) is the ultimate test of its usefulness.
2. Expert Physician Review: Qualitative review is crucial for understanding the clinical context and significance behind statistical differences (e.g., recognizing under-segmentation vs. clinically useful over-segmentation).
3. Radiomics feature stability: Analyzing how segmentation variations impact quantitative features is essential for understanding the reliability of any derived predictive models.

Section 7 — Clinical Adoption & Workflow Impact

7.1 Usability — How easy is it to interpret and use the AI mask compared to GT in clinical workflow?

0 ☐ 1 ■ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = much harder, 5 = much easier)

Reason: Although the mask is simply shaped, its severe under-segmentation makes it much harder to use. A clinician can't just refine the boundary; they would need to redraw most of the tumor from scratch, defeating the purpose of the tool and complicating the workflow.

7.2 Clinical actionability — Would the AI–GT differences change clinical decisions (for example, radiotherapy target, surgery extent, follow-up planning)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ■ (0 = no impact, 5 = strong impact)

Reason: The differences could have a strong, dangerously negative impact. Using this mask for radiotherapy planning might result in significant under-treatment of the tumor, and for surgery, it could lead to an inadequate resection margin. Clinical decisions based on this mask would be fundamentally flawed.

7.3 Robustness and variability — How consistent and reliable are AI masks across different cases and observers?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ■ (0 = very inconsistent, 5 = very consistent)

Reason: The model consistently fails by producing a severely under-segmented mask in all reviewed cases, showing a systematic and reliable flaw in its approach.

7.4 Time-saving potential — How much time could this AI save compared to manual delineation?

0 ■ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no time saved, 5 = saves most effort)

Reason: This AI would not save any time. The generated mask is clinically unusable and requires a physician to completely redelineate it. It adds no value to the clinical workflow and might even waste time by offering a misleading starting point.

7.5 Trust in clinical use — How confident would you be using AI masks for high-stakes tasks (e.g., radiotherapy planning) without physician refinement?

0 ■ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no trust, 5 = complete trust)

Reason: There is no trust in using this mask for any clinical task. Its fundamental inability to accurately capture the true tumor volume makes it dangerously unreliable for high-stakes applications such as treatment planning.

7.6 Confidence maps/interpretability (if available) — Do uncertainty or confidence maps improve your trust in AI outputs?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no improvement, 5 = very strong improvement)

Score: N/A

Reason: Confidence maps were not available. However, even if they had been, they would not foster trust in a model that so fundamentally fails at the primary segmentation task. Interpretability cannot fix a completely incorrect output.

7.7 Role of AI in clinical practice — In your opinion, AI-based segmentation should:

☐ Replace physicians entirely  
■ Provide an initial mask for physicians to refine  
☐ Assist physicians as a support/check tool

Reason: While the ideal role for AI is to serve as a reliable aid for physician refinement, this specific model (3DMedSam) does not even meet that standard. Its output is not a suitable starting point and therefore has no useful role in clinical practice.

**End**

**Comparative Clinical Assessment of Deep Learning Tumor Masks vs. Ground Truth in Lung Cancer**

Model under review: Reconnet

**Segmentation Results**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Analysis Mode | Model | Dice Score | IoU Score | Hausdorff Distance |
| Test |  |  |  |  |
|  |  |  |  |
| **Reconnet** | 0.79 0.11 | 0.67 0.13 | 11.39 3.96 |

**Prediction Result:**

Radiomics extracted from Ground Truth used for prediction tasks

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Validation | | | | | | Test | | | | | | | Names | | | |
| Acc | Precision | Recall | F1 | AUC | Specificity | Acc | Precision | Recall | F1 | AUC | Specificity | FSA/AEA | | CA | Model | SL/SSL |
| 0.88 | 0.78 | 0.88 | 0.83 | 0.54 | 1 | 0.87 | 0.76 | 0.87 | 0.81 | 0.71 | 1 | Mutual\_Info\_Gain\_Ratio | | ExtraTree | GT | SSL |
| 0.74 | 0.67 | 0.74 | 0.66 | 0.55 | 0.98 | 0.87 | 0.79 | 0.87 | 0.83 | 0.63 | 0.99 | SL |

Radiomics extracted from DL Reconnet used for prediction tasks

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Validation | | | | | | Test | | | | | | Names | | | |
| Acc | Precision | Recall | F1 | AUC | Specificity | Acc | Precision | Recall | F1 | AUC | Specificity | FSA/AEA | CA | Model | SL/SSL |
| 0.88 | 0.78 | 0.88 | 0.83 | 0.65 | 1 | 0.86 | 0.75 | 0.86 | 0.80 | 0.71 | 1 | UMAP | AdaBoost | Reconnet | SSL |
| 0.75 | 0.56 | 0.75 | 0.64 | 0.50 | 1 | 0.88 | 0.77 | 0.88 | 0.82 | 0.5 | 1 | SL |

**Start**

Section 1 — Overall Impression

1.1 Does the AI mask define a tumor volume that is more clinically meaningful than the Ground Truth (GT) for diagnosis and visual interpretation?

0 ☐ 1 ☐ 2 ☐ 3 ■ 4 ☐ 5 ☐ (0 = not at all, 5 = very strongly yes)

Reason: The AI mask is roughly as clinically meaningful as the GT; they are essentially equivalent. The visual analysis reveals a high level of agreement between the AI and GT masks, indicating that the model accurately reproduces the manual delineation standard. This is further supported by its identical performance in the survival prediction task (AUC 0.71), as it offers no new or different information compared to the GT.

Section 2 — Spatial Differences

2.1 Are the AI–GT differences concentrated in the peritumoral zone (≈5–15 mm around the tumor)?

0 ☐ 1 ☐ 2 ■ 3 ☐ 4 ☐ 5 ☐ (0 = not at all, 5 = strongly concentrated)

Reason: The small differences between the AI and GT masks are not consistently found in the peritumoral zone. Instead, they show as minor, localized disagreements along the shared tumor boundary, aligning with slight inter-observer variability rather than intentional inclusion of surrounding tissue.

2.2 How clinically meaningful are the regions where the AI mask differs from GT (including over-segmentation or under-segmentation, such as peritumoral expansion, missed necrotic cores, spiculated margins)? Or, how clinically justified are the AI–GT differences (whether due to additional regions or excluded regions)?

0 ☐ 1 ■ 2 ☐ 3 ☐ 4 ☐ 5 ☐

Reason: The regions of difference are not clinically significant. They exhibit minor boundary fluctuations that do not alter or add significant biological information. The deviations are essentially insignificant and do not change the interpretation of the tumor's volume or extent.

Section 3 — Clinical Interpretation of Differences

3.1 How well does the AI mask preserve internal tumor heterogeneity (for example: solid vs. ground-glass opacity, necrotic vs. viable regions)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ■

Reason: The AI mask maintains internal heterogeneity as effectively as the Ground Truth because it accurately outlines the same tumor volume. It successfully captures the variations within the lesion without missing or adding external, non-tumorous regions.

3.2 Boundary quality: compared to GT, how noisy or fragmented are AI boundaries?

0 ☐ 1 ■ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = much less noisy, 5 = much more noisy)

Reason: The AI-generated boundaries are high quality. They are smooth and continuous, without noise or fragmentation. The minor deviations from the GT are just small shifts within a coherent line, not a sign of an unstable or unpredictable segmentation process.

Section 4 — Link to Survival Prediction

4.1 Do the AI–GT differences likely capture prognostically relevant tissue that improves survival prediction?

0 ■ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

Reason: No, the differences between AI-GT are minor and clinically insignificant, similar to slight inter-observer variability rather than the systematic detection of new biological information. This is supported by the identical survival prediction performance (AUC 0.71) for both the AI and GT masks, indicating that no additional prognostically relevant tissue was identified.

4.2 How would you rate the effect of the AI’s deviation from GT (larger or smaller volume) on survival prediction?

0 ☐ 1 ☐ 2 ■ 3 ☐ 4 ☐ 5 ☐ (0 = strongly harmful, 5 = strongly beneficial)

Reason: The AI's deviation from the GT is minimal and has a neutral effect on survival prediction. The model accurately replicates the manual segmentation, resulting in the same prognostic outcome as the GT. It is neither harmful nor beneficial but simply equivalent.

Section 5 — Trust & Preference (Survival Prediction Focus)

5.1 How likely is it that AI-generated masks are more accurate than GT in some cases (for example, subtle lesions missed by physicians, or over-delineation that reduces noise) in survival prediction tasks?

0 ☐ 1 ■ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = very unlikely, 5 = very likely)

Reason: It is doubtful that the AI masks are more accurate. The quantitative data show their prognostic performance is identical to the Ground Truth (AUC 0.71). The visual analysis confirms the masks are essentially geometric replicas of the GT, meaning they contain the same clinical information and therefore offer no additional accuracy.

5.2 Would you prefer to use the AI mask (rather than GT) as input for survival prediction?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ■ 5 ☐ (0 = not at all, 5 = strongly yes)

Reason: Yes, I prefer using the AI mask. Although it doesn't improve prognostic accuracy, it performs just as well as manual delineation automatically. This offers significant workflow efficiency and time savings without compromising prediction quality, making it a more practical choice.

Section 6 — Summaries and Overall Assessment

6.1 Diagnosis-oriented evaluation — How successful is the AI mask in lesion delineation compared to GT?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ■ (0 = complete failure, 5 = highly successful)

Reason: The AI mask is highly successful in replicating the Ground Truth. The visual analysis reveals a high degree of geometric concordance, indicating that the model has effectively learned to reproduce the manual delineation standard for diagnostic purposes.

6.2 Prediction-oriented evaluation — How successful is the AI mask in contributing to survival prediction compared to GT?

0 ☐ 1 ☐ 2 ☐ 3 ■ 4 ☐ 5 ☐ (0 = no added value, 5 = very strong added value)

Reason: The AI mask successfully replicates the prognostic performance of the GT (both with an AUC of 0.71), but it offers no added value. It contributes by reliably automating the standard, but it does not enhance the prediction by capturing new biological information.

6.3 Overall confidence in your assessment

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ■ (0 = no confidence, 5 = complete confidence)

Reason: Confidence is complete because the visual evidence of high geometric similarity aligns perfectly with the quantitative data (strong Dice/IoU scores and identical AUC), creating a consistent and unambiguous profile for this model.

6.4 Do you think geometric evaluation metrics (e.g., Dice, IoU, Hausdorff distance) are sufficient to establish clinical trust in AI segmentation?

0 ☐ 1 ■ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not sufficient at all, 5 = fully sufficient)

Reason: No. While the geometric metrics for this model accurately reflected its clinical equivalence to GT, our broader evaluation (including the 3DAttention and 3DMedSam models) proves they are insufficient. A model can have excellent or poor geometric scores that are completely disconnected from its actual clinical utility for a downstream task.

6.5 What additional types of evaluation do you think are necessary for trustworthy AI segmentation (e.g., radiomics stability, survival prediction, physician review, workflow impact, and others), and why do you think those are effective in building trustworthiness?

Reason: Reliable AI evaluation requires a multi-modal approach. It requires:

1. Downstream Task Performance: The key evaluation is measuring the AI's effect on a specific clinical outcome, like survival prediction, to determine its true value.
2. Expert Physician Review: Conducting a qualitative analysis is crucial to understanding the clinical significance of any AI-GT discrepancies.
3. Radiomics Feature Stability: This is necessary to ensure that the quantitative features obtained from the segmentations are consistent and dependable.

Section 7 — Clinical Adoption & Workflow Impact

7.1 Usability — How easy is it to interpret and use the AI mask compared to GT in clinical workflow?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ■ (0 = much harder, 5 = much easier)

Reason: The AI mask is much easier to use because it provides an accurate, automated delineation that conforms to the clinical standard. A physician can review and make minor edits, which is much quicker and easier than creating a mask from scratch.

7.2 Clinical actionability — Would the AI–GT differences change clinical decisions (for example, radiotherapy target, surgery extent, follow-up planning)?

0 ■ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no impact, 5 = strong impact)

Reason: The differences are clinically insignificant and would not influence decisions about radiotherapy, surgery, or follow-up. The AI mask offers the same anatomical information as the manual Ground Truth.

7.3 Robustness and variability — How consistent and reliable are AI masks across different cases and observers?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ■ (0 = very inconsistent, 5 = very consistent)

Reason: The model shows high consistency across all evaluated cases, reliably producing segmentations very close to the Ground Truth. This indicates it is a dependable tool that can be trusted to replicate the manual standard.

7.4 Time-saving potential — How much time could this AI save compared to manual delineation?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ■ (0 = no time saved, 5 = saves most effort)

Reason: The AI can save most of the time and effort needed for manual delineation. It cuts down the physician's work from hours of contouring to minutes of reviewing and refining.

7.5 Trust in clinical use — How confident would you be using AI masks for high-stakes tasks (e.g., radiotherapy planning) without physician refinement?

0 ☐ 1 ☐ 2 ☐ 3 ■ 4 ☐ 5 ☐ (0 = no trust, 5 = complete trust)

Reason: Although trust is high, I would still need a physician to refine decisions for high-stakes tasks, such as radiotherapy planning. The AI offers an excellent and reliable initial point, but the final clinical responsibility requires an expert check and sign-off.

7.6 Confidence maps/interpretability (if available) — Do uncertainty or confidence maps improve your trust in AI outputs?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no improvement, 5 = very strong improvement)

Score: N/A

Reason: Confidence maps were not available for this evaluation. However, if they were, they would likely enhance trust by highlighting areas where the model had lower certainty, enabling a physician to focus review on those specific regions.

7.7 Role of AI in clinical practice — In your opinion, AI-based segmentation should:

☐ Replace physicians entirely  
■ Provide an initial mask for physicians to refine  
☐ Assist physicians as a support/check tool

Reason: This model exemplifies an AI tool designed to act as a highly efficient assistant. Its primary role is to generate a high-quality initial segmentation that significantly speeds up the clinical workflow, which the physician then quickly reviews and completes.

**End**

**Comparative Clinical Assessment of Deep Learning Tumor Masks vs. Ground Truth in Lung Cancer**

Model under review: ResUnet

**Segmentation Results**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Analysis Mode | Model | Dice Score | IoU Score | Hausdorff Distance |
| Test |  |  |  |  |
| **ResUNet** | 0.75 0.09 | 0.61 0.1 | 19.91 6.91 |

**Prediction Result:**

Radiomics extracted from Ground Truth used for prediction tasks

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Validation | | | | | | Test | | | | | | | Names | | | |
| Acc | Precision | Recall | F1 | AUC | Specificity | Acc | Precision | Recall | F1 | AUC | Specificity | FSA/AEA | | CA | Model | SL/SSL |
| 0.88 | 0.78 | 0.88 | 0.83 | 0.54 | 1 | 0.87 | 0.76 | 0.87 | 0.81 | 0.71 | 1 | Mutual\_Info\_Gain\_Ratio | | ExtraTree | GT | SSL |
| 0.74 | 0.67 | 0.74 | 0.66 | 0.55 | 0.98 | 0.87 | 0.79 | 0.87 | 0.83 | 0.63 | 0.99 | SL |

Radiomics extracted from DL ResUNet used for prediction tasks

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Validation | | | | | | Test | | | | | | | Names | | | |
| Acc | Precision | Recall | F1 | AUC | Specificity | Acc | Precision | Recall | F1 | AUC | Specificity | FSA/AEA | | CA | Model | SL/SSL |
| 0.88 | 0.78 | 0.88 | 0.83 | 0.30 | 1 | 0.84 | 0.71 | 0.84 | 0.77 | 0.67 | 1 | Mutual Information | | MLP | ResUNet | SSL |
| 0.75 | 0.56 | 0.75 | 0.64 | 0.56 | 1 | 0.83 | 0.69 | 0.83 | 0.75 | 0.68 | 1 | SL |

**Start**

Section 1 — Overall Impression

1.1 Does the AI mask define a tumor volume that is more clinically meaningful than the Ground Truth (GT) for diagnosis and visual interpretation?

0 ☐ 1 ☐ 2 ■ 3 ☐ 4 ☐ 5 ☐ (0 = not at all, 5 = very strongly yes)

Reason: The AI mask is less clinically meaningful than the GT because of inconsistent boundary delineation. The model struggles to accurately capture complex or spiculated margins, leading to a noisy and imprecise mask. This geometric instability likely causes errors in the radiomics features, which explains the model's poorer performance in the survival prediction task.

Section 2 — Spatial Differences

2.1 Are the AI–GT differences concentrated in the peritumoral zone (≈5–15 mm around the tumor)?

0 ☐ 1 ■ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not at all, 5 = strongly concentrated)

Reason: The AI-GT differences are not a systematic concentration in the peritumoral zone. Instead, they manifest as random, localized errors directly on the tumor boundary, where the model's delineation appears noisy and struggles to conform to the true margin.

2.2 How clinically meaningful are the regions where the AI mask differs from GT (including over-segmentation or under-segmentation, such as peritumoral expansion, missed necrotic cores, spiculated margins)? Or, how clinically justified are the AI–GT differences (whether due to additional regions or excluded regions)?

0 ■ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

Reason: The regions of difference are not clinically significant. They indicate noise and instability in the segmentation, especially around complex margins, rather than representing a specific, justifiable biological area. These errors are not clinically justified and weaken the quality of the mask for prognostic analysis.

Section 3 — Clinical Interpretation of Differences

3.1 How well does the AI mask preserve internal tumor heterogeneity (for example: solid vs. ground-glass opacity, necrotic vs. viable regions)?

0 ☐ 1 ☐ 2 ■ 3 ☐ 4 ☐ 5 ☐

Reason: The model's inconsistent boundaries prevent it from consistently capturing the full internal heterogeneity. By being imprecise at the edges, it either overlooks peripheral textures or mistakenly includes nearby non-tumorous tissue, resulting in an unreliable depiction of the lesion's complete internal complexity.

3.2 Boundary quality: compared to GT, how noisy or fragmented are AI boundaries?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ■ 5 ☐ (0 = much less noisy, 5 = much more noisy)

Reason: The AI boundaries are much noisier and less consistent than the smooth, clean Ground Truth contours. This visual assessment is strongly supported by the high Hausdorff distance (19.91), which clearly confirms the presence of large, erratic boundary errors.

Section 4 — Link to Survival Prediction

4.1 Do the AI–GT differences likely capture prognostically relevant tissue that improves survival prediction?

0 ■ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

Reason: No. The differences between the AI and GT masks are not a systematic capture of a specific biological region but rather random, noisy errors at the tumor boundary. This introduction of noise, rather than new information, is likely the cause of the model's inferior performance (AUC 0.67) compared to the Ground Truth (AUC 0.71).

4.2 How would you rate the effect of the AI’s deviation from GT (larger or smaller volume) on survival prediction?

0 ☐ 1 ■ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = strongly harmful, 5 = strongly beneficial)

Reason: The AI's deviation harms survival prediction. The inconsistent and noisy boundaries generate an unstable tumor representation, which reduces the quality and reliability of the extracted radiomics features, resulting in a less accurate prognostic model.

Section 5 — Trust & Preference (Survival Prediction Focus)

5.1 How likely is it that AI-generated masks are more accurate than GT in some cases (for example, subtle lesions missed by physicians, or over-delineation that reduces noise) in survival prediction tasks?

0 ■ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = very unlikely, 5 = very likely)

Reason: This AI mask probably won't be more accurate in predicting survival. The data shows it performs worse than the Ground Truth (AUC 0.67 vs. 0.71). Also, the visual analysis confirms that its deviations from the GT are due to random boundary noise, not meaningful biological information.

5.2 Would you prefer to use the AI mask (rather than GT) as input for survival prediction?

0 ■ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not at all, 5 = strongly yes)

Reason: I would not depend on this AI mask for survival predictions. The Ground Truth offers a more stable and reliable segmentation, resulting in a more accurate prognostic model. Using the noisy and inconsistent AI mask would lower the prediction quality.

Section 6 — Summaries and Overall Assessment

6.1 Diagnosis-oriented evaluation — How successful is the AI mask in lesion delineation compared to GT?

0 ☐ 1 ☐ 2 ■ 3 ☐ 4 ☐ 5 ☐ (0 = complete failure, 5 = highly successful)

Reason: The model is only partially effective. It often fails to accurately outline complex and spiculated margins, resulting in noisy and geometrically unstable boundaries that are less reliable than the Ground Truth for diagnostic purposes.

6.2 Prediction-oriented evaluation — How successful is the AI mask in contributing to survival prediction compared to GT?

0 ☐ 1 ■ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no added value, 5 = very strong added value)

Reason: The AI mask impairs survival prediction. Noisy and inconsistent segmentations lead to errors in the radiomics features extracted, resulting in a model with lower prognostic accuracy (AUC 0.67) compared to the more stable Ground Truth masks (AUC 0.71).

6.3 Overall confidence in your assessment

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ■ (0 = no confidence, 5 = complete confidence)

Reason: Confidence is high because the visual observation of noisy, inconsistent boundaries clearly explains the poor quantitative results (low Dice, high Hausdorff, and low AUC). The evidence remains consistent across both evaluation methods.

6.4 Do you think geometric evaluation metrics (e.g., Dice, IoU, Hausdorff distance) are sufficient to establish clinical trust in AI segmentation?

0 ■ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not sufficient at all, 5 = fully sufficient)

Reason: No, they are definitely not sufficient. This entire comparative analysis shows that geometric scores can be misleading. A model with low geometric scores (3DAttention) may still have high clinical value, while a model with higher scores (ResUnet) might be less useful because of the type of errors it makes—random noise versus systematic, meaningful deviations.

6.5 What additional types of evaluation do you think are necessary for trustworthy AI segmentation (e.g., radiomics stability, survival prediction, physician review, workflow impact, and others), and why do you think those are effective in building trustworthiness?

Reason: To build trust, evaluation must extend beyond geometry and include:

* 1. Downstream Task Impact: Evaluating performance on the final clinical goal (e.g., survival prediction) is the most critical measure of value.
  2. Qualitative Physician Review: Expert assessment is crucial to understand why a model succeeds or fails, offering clinical context to the data.
  3. Feature Stability Analysis: This is necessary to ensure that the quantitative features used for prediction remain reliable despite small variations in segmentation.

Section 7 — Clinical Adoption & Workflow Impact

7.1 Usability — How easy is it to interpret and use the AI mask compared to GT in clinical workflow?

0 ☐ 1 ■ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = much harder, 5 = much easier)

Reason: The mask is much harder to use in a clinical workflow. Its noisy and inconsistent boundaries require extensive, time-consuming corrections by a physician, rather than just simple refinements. In many cases, it is easier to start from scratch.

7.2 Clinical actionability — Would the AI–GT differences change clinical decisions (for example, radiotherapy target, surgery extent, follow-up planning)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ■ 5 ☐ (0 = no impact, 5 = strong impact)

Reason: The differences could significantly negatively affect outcomes. Noisy boundaries result in an unreliable and inaccurate tumor volume, which might cause incorrect staging, flawed radiotherapy planning (such as missing the tumor or irradiating healthy tissue), or improper surgical margins.

7.3 Robustness and variability — How consistent and reliable are AI masks across different cases and observers?

0 ☐ 1 ■ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = very inconsistent, 5 = very consistent)

Reason: The model is very inconsistent. Its performance is unreliable, especially on tumors with complex or spiculated margins, where it often fails to accurately define the boundaries. This lack of robustness makes it untrustworthy for clinical use.

7.4 Time-saving potential — How much time could this AI save compared to manual delineation?

0 ☐ 1 ■ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no time saved, 5 = saves most effort)

Reason: This AI would save almost no time. The poor quality of the initial mask would require extensive editing or even complete redrawing by the physician, largely negating any potential workflow efficiencies.

7.5 Trust in clinical use — How confident would you be using AI masks for high-stakes tasks (e.g., radiotherapy planning) without physician refinement?

0 ■ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no trust, 5 = complete trust)

Reason: This mask cannot be trusted for high-stakes tasks without full physician supervision and significant refinement. Its inherent instability and unreliability make it unsafe for direct clinical use in applications like treatment planning.

7.6 Confidence maps/interpretability (if available) — Do uncertainty or confidence maps improve your trust in AI outputs?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no improvement, 5 = very strong improvement)  
Score: N/A

Reason: Confidence maps weren't available. If they had been, they could have helped by clearly marking noisy and uncertain boundary areas, but they wouldn't have fixed the model's fundamental poor performance.

7.7 Role of AI in clinical practice — In your opinion, AI-based segmentation should:

☐ Replace physicians entirely  
■ Provide an initial mask for physicians to refine  
☐ Assist physicians as a support/check tool

Reason: While the ideal role for this type of AI is to serve as a starting point for refinement, this specific model (ResUnet) often fails to produce a mask of sufficient quality to be useful even in that assistive capacity. It does not meet the standard required for a reliable clinical tool.

**End**

**Comparative Clinical Assessment of Deep Learning Tumor Masks vs. Ground Truth in Lung Cancer**

Model under review: VNet

**Segmentation Results**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Analysis Mode | Model | Dice Score | IoU Score | Hausdorff Distance |
| Test | **VNet** | 0.83 0.07 | 0.71 0.09 | 10.04 3.73 |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

**Prediction Result:**

Radiomics extracted from Ground Truth used for prediction tasks

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Validation | | | | | | Test | | | | | | | Names | | | |
| Acc | Precision | Recall | F1 | AUC | Specificity | Acc | Precision | Recall | F1 | AUC | Specificity | FSA/AEA | | CA | Model | SL/SSL |
| 0.88 | 0.78 | 0.88 | 0.83 | 0.54 | 1 | 0.87 | 0.76 | 0.87 | 0.81 | 0.71 | 1 | Mutual\_Info\_Gain\_Ratio | | ExtraTree | GT | SSL |
| 0.74 | 0.67 | 0.74 | 0.66 | 0.55 | 0.98 | 0.87 | 0.79 | 0.87 | 0.83 | 0.63 | 0.99 | SL |

Radiomics extracted from Vnet used for prediction tasks

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Validation | | | | | | Test | | | | | | Names | | | |
| Acc | Precision | Recall | F1 | AUC | Specificity | Acc | Precision | Recall | F1 | AUC | Specificity | FSA/AEA | CA | Model | SL/SSL |
| 0.88 | 0.78 | 0.88 | 0.83 | 0.65 | 1 | 0.88 | 0.77 | 0.88 | 0.82 | 0.76 | 1 | LASSO | LGBM | VNet | SSL |
| 0.74 | 0.55 | 0.74 | 0.63 | 0.58 | 0.99 | 0.87 | 0.77 | 0.87 | 0.81 | 0.40 | 0.99 | SL |

**Start**

Section 1 — Overall Impression

1.1 Does the AI mask define a tumor volume that is more clinically meaningful than the Ground Truth (GT) for diagnosis and visual interpretation?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ■ (0 = not at all, 5 = very strongly yes)

Reason: The AI mask is more clinically meaningful because it strikes an optimal balance between geometric accuracy and capturing prognostically relevant information. While maintaining high fidelity to the Ground Truth (as shown by the best Dice/Hausdorff scores), it makes subtle, intelligent expansions into the peritumoral zone. This "smart" over-segmentation allows it to outperform the GT in survival prediction (AUC 0.76 vs. 0.71) without sacrificing diagnostic precision.

Section 2 — Spatial Differences

2.1 Are the AI–GT differences concentrated in the peritumoral zone (≈5–15 mm around the tumor)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ■ 5 ☐ (0 = not at all, 5 = strongly concentrated)

Reason: Yes, the minor differences are mostly focused in the peritumoral zone. Unlike the broad over-segmentation of 3DAttention, VNet makes subtle and precise expansions at the tumor boundary, indicating a more refined method of capturing the immediate tumor microenvironment without large geometric mistakes.

2.2 How clinically meaningful are the regions where the AI mask differs from GT (including over-segmentation or under-segmentation, such as peritumoral expansion, missed necrotic cores, spiculated margins)? Or, how clinically justified are the AI–GT differences (whether due to additional regions or excluded regions)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ■

Reason: The regions of difference are highly clinically significant. These subtle, justified expansions into the peritumoral area likely provide additional prognostic information that enables the VNet model to outperform the Ground Truth in survival prediction (AUC 0.76 vs. 0.71). The model's "corrections" to the GT are both minor and clinically meaningful.

Section 3 — Clinical Interpretation of Differences

3.1 How well does the AI mask preserve internal tumor heterogeneity (for example: solid vs. ground-glass opacity, necrotic vs. viable regions)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ■

Reason: The AI mask performs exceptionally well in this area. It not only maintains the internal heterogeneity of the solid tumor core with high accuracy to the GT, but it also improves it by subtly including nearby peritumoral textures (like GGO), offering a more comprehensive biological view of the lesion.

3.2 Boundary quality: compared to GT, how noisy or fragmented are AI boundaries?

0 ■ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = much less noisy, 5 = much more noisy)

Reason: The AI boundaries are of top quality—smooth, continuous, and free of noise or fragmentation. This visual assessment is strongly supported by the quantitative data, as VNet achieved the best (lowest) Hausdorff distance score among all models, indicating superior boundary accuracy.

Section 4 — Link to Survival Prediction

4.1 Do the AI–GT differences likely capture prognostically relevant tissue that improves survival prediction?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ■

Reason: Yes, absolutely. This is supported by the quantitative data, which shows a better survival prediction (AUC 0.76) for the VNet mask compared to the GT (AUC 0.71). The visual analysis indicates that the differences between AI and GT are subtle, involving targeted expansions into the peritumoral zone, directly suggesting that the model detects prognostically important tissue that the standard manual delineation misses.

4.2 How would you rate the effect of the AI’s deviation from GT (larger or smaller volume) on survival prediction?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ■ (0 = strongly harmful, 5 = strongly beneficial)

Reason: The AI's deviation is highly beneficial. By making small, smart expansions to develop a more complete biological target volume, the model provides a richer, more accurate input for the prognostic model. This results in a clearly improved prediction of patient outcomes, confirming the clinical value of its deviation.

Section 5 — Trust & Preference (Survival Prediction Focus)

5.1 How likely is it that AI-generated masks are more accurate than GT in some cases (for example, subtle lesions missed by physicians, or over-delineation that reduces noise) in survival prediction tasks?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ■ (0 = very unlikely, 5 = very likely)

Reason: It is very likely, as demonstrated by the quantitative results. The VNet model achieved a higher AUC (0.76) for survival prediction than the Ground Truth (0.71). Our qualitative analysis confirmed that this is because of its ability to make subtle, clinically relevant expansions into the peritumoral zone, thereby capturing more comprehensive prognostic information.

5.2 Would you prefer to use the AI mask (rather than GT) as input for survival prediction?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ■ (0 = not at all, 5 = strongly yes)

Reason: Yes, for the specific task of survival prediction, the VNet mask is highly preferred. It has been shown through empirical evidence to produce a more accurate prognostic model. Its ability to balance geometric precision with the capture of a broader biological target volume makes it the better choice for this purpose.

Section 6 — Summaries and Overall Assessment

6.1 Diagnosis-oriented evaluation — How successful is the AI mask in lesion delineation compared to GT?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ■ (0 = complete failure, 5 = highly successful)

Reason: The AI mask is highly effective and represents the latest advancement in lesion delineation. It achieved the highest geometric concordance scores (Dice, IoU, and Hausdorff distance) among all models, demonstrating its superior ability to accurately and reliably replicate the manual standard.

6.2 Prediction-oriented evaluation — How successful is the AI mask in contributing to survival prediction compared to GT?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ■ (0 = no added value, 5 = very strong added value)

Reason: The model offers significantly increased value, surpassing the Ground Truth's prognostic ability (AUC 0.76 vs. 0.71). It accomplishes this by making subtle yet clinically important expansions into the peritumoral zone, thereby capturing a more comprehensive and prognostically relevant biological volume.

6.3 Overall confidence in your assessment

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ■ (0 = no confidence, 5 = complete confidence)

Reason: Confidence is absolute because this is the only model where superior quantitative performance across all metrics was perfectly matched and supported by the qualitative visual review. The evidence for its excellence is consistent, robust, and clear.

6.4 Do you think geometric evaluation metrics (e.g., Dice, IoU, Hausdorff distance) are sufficient to establish clinical trust in AI segmentation?

0 ■ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not sufficient at all, 5 = fully sufficient)

No, they are completely inadequate. This thorough five-model review has clearly shown that geometric scores alone are poor indicators of clinical usefulness. A model's true value in the real world depends on how clinically relevant its deviations from the ground truth are, not just how large those deviations are.

6.5 What additional types of evaluation do you think are necessary for trustworthy AI segmentation (e.g., radiomics stability, survival prediction, physician review, workflow impact, and others), and why do you think those are effective in building trustworthiness?

Reason: Building trust in AI needs a comprehensive evaluation framework that goes beyond just geometry. The most important parts, as shown by this exercise, are:

1. Downstream Task Performance: This serves as the ultimate indicator of clinical usefulness, evaluating how the segmentation affects real-world outcomes such as survival prediction.
2. Qualitative Expert Review: This is crucial for providing the clinical context and interpretation of why a model succeeds or fails, turning numbers into actionable insights.
3. Radiomics feature stability: This is essential to ensure that the quantitative data derived from the masks is reliable and not overly sensitive to small segmentation variations.

Section 7 — Clinical Adoption & Workflow Impact

7.1 Usability — How easy is it to interpret and use the AI mask compared to GT in clinical workflow?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ■ (0 = much harder, 5 = much easier)

Reason: The AI mask is much easier to use. It offers an exceptionally accurate and reliable starting point that requires little to no physician refinement. Its ability to intelligently capture prognostically relevant areas also makes it simpler to interpret from a clinical decision-making perspective.

7.2 Clinical actionability — Would the AI–GT differences change clinical decisions (for example, radiotherapy target, surgery extent, follow-up planning)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ■ 5 ☐ (0 = no impact, 5 = strong impact)

Reason: The differences could have a significant, positive effect. The subtle, intelligent inclusion of the peritumoral zone might lead to more informed clinical decisions, such as adjusting radiotherapy target volumes (e.g., CTV) or recommending more aggressive follow-up for patients whose AI mask indicates a more invasive phenotype.

7.3 Robustness and variability — How consistent and reliable are AI masks across different cases and observers?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ■ (0 = very inconsistent, 5 = very consistent)

Reason: The model is highly consistent and reliable. It delivered high-quality, accurate segmentations across all reviewed cases, showing strong robustness even with different tumor shapes and complexities.

7.4 Time-saving potential — How much time could this AI save compared to manual delineation?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ■ (0 = no time saved, 5 = saves most effort)

Reason: This AI would save the most effort. Its high accuracy shifts the physician's role from manual contouring to quick review and verification, improving workflow efficiency.

7.5 Trust in clinical use — How confident would you be using AI masks for high-stakes tasks (e.g., radiotherapy planning) without physician refinement?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ■ 5 ☐ (0 = no trust, 5 = complete trust)

Reason: Confidence in this model is very high, far exceeding that of others. While final physician approval is always necessary for high-stakes tasks, the model's consistency and accuracy suggest it could be used with only minimal verification, nearly enabling automatic segmentation.

7.6 Confidence maps/interpretability (if available) — Do uncertainty or confidence maps improve your trust in AI outputs?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no improvement, 5 = very strong improvement)

Score: N/A

Reason: Confidence maps were not available. However, for a model this accurate, such maps would be extremely useful for establishing the trust needed for full clinical use, especially by showing the subtle peritumoral areas where it differs from a standard GT.

7.7 Role of AI in clinical practice — In your opinion, AI-based segmentation should:

☐ Replace physicians entirely  
■ Provide an initial mask for physicians to refine  
☐ Assist physicians as a support/check tool

Reason: This model perfectly represents the ideal role of AI in current clinical practice. It functions as a highly dependable and intelligent assistant that offers a near-perfect initial mask, which a physician can quickly verify, adjust if needed, and approve, thereby improving both efficiency and the quality of clinical data.

**End**

**D2**

**Comparative Clinical Assessment of Deep Learning Tumor Masks vs. Ground Truth in Lung Cancer**

Model under review: 3DAttention

**Segmentation Results**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Analysis Mode | Model | Dice Score | IoU Score | Hausdorff Distance |
| Test |  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
| **3DAttention** | 0.83 0.04 | 0.72 0.04 | 12.12 3.73 |

**Prediction Result:**

Radiomics extracted from Ground Truth used for prediction tasks

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Validation | | | | | | Test | | | | | | | Names | | | |
| Acc | Precision | Recall | F1 | AUC | Specificity | Acc | Precision | Recall | F1 | AUC | Specificity | FSA/AEA | | CA | Model | SL/SSL |
| 0.88 | 0.78 | 0.88 | 0.83 | 0.54 | 1 | 0.87 | 0.76 | 0.87 | 0.81 | 0.71 | 1 | Mutual\_Info\_Gain\_Ratio | | ExtraTree | GT | SSL |
| 0.74 | 0.67 | 0.74 | 0.66 | 0.55 | 0.98 | 0.87 | 0.79 | 0.87 | 0.83 | 0.63 | 0.99 | SL |

Radiomics extracted from DL 3D attention used for prediction tasks

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Validation | | | | | | Test | | | | | | | Names | | | |
| Acc | Precision | Recall | F1 | AUC | Specificity | Acc | Precision | Recall | F1 | AUC | Specificity | FSA/AEA | | CA | Model | SL/SSL |
| 0.79 | 0.80 | 0.79 | 0.80 | 0.69 | 0.87 | 0.81 | 0.83 | 0.81 | 0.82 | 0.76 | 0.88 | Guassian Random Projection | | Voting Classifier | 3DAtt | SSL |
| 0.61 | 0.64 | 0.61 | 0.62 | 0.53 | 0.69 | 0.72 | 0.79 | 0.72 | 0.75 | 0.55 | 0.79 | SL |

**Start**

Section 1 — Overall Impression

1.1 Does the AI mask define a tumor volume that is more clinically meaningful than the Ground Truth (GT) for diagnosis and visual interpretation?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not at all, 5 = very strongly yes)

Reason:

The AI mask does not provide a clinically more meaningful tumor volume definition than the ground truth (GT) for diagnosis and visual interpretation, especially in the case of irregular, spiculated borders and necrotic regions of tumors.

Section 2 — Spatial Differences

2.1 Are the AI–GT differences concentrated in the peritumoral zone (≈5–15 mm around the tumor)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not at all, 5 = strongly concentrated)

Reason:

More data is needed to answer this question correctly, such as information about the size of the image slices. In general, however, the answer is no, especially for spiculated lesions. Additionally, in some cases, major misregistration between the tumor and AI masks is apparent.

2.2 How clinically meaningful are the regions where the AI mask differs from GT (including over-segmentation or under-segmentation, such as peritumoral expansion, missed necrotic cores, spiculated margins)? Or, how clinically justified are the AI–GT differences (whether due to additional regions or excluded regions)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

Reason:

This difference is clinically significant, with a greater impact observed in peritumoral and spiculated margins compared to necrotic areas.

Section 3 — Clinical Interpretation of Differences

3.1 How well does the AI mask preserve internal tumor heterogeneity (for example: solid vs. ground-glass opacity, necrotic vs. viable regions)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

Reason:

AI masks fail to preserve internal tumor heterogeneity, resulting in an overestimation of tumor volume in ground-glass opacity and necrotic regions.

3.2 Boundary quality: compared to GT, how noisy or fragmented are AI boundaries?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = much less noisy, 5 = much more noisy)

Reason:

Boundary is more noisy in AI masks compared to GT masks.

Section 4 — Link to Survival Prediction

4.1 Do the AI–GT differences likely capture prognostically relevant tissue that improves survival prediction?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

Reason:

The answer to this question requires multiple experiments and statistical analysis.

4.2 How would you rate the effect of the AI’s deviation from GT (larger or smaller volume) on survival prediction?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = strongly harmful, 5 = strongly beneficial)

Reason:

Depending on the tumor's location and the type of lesions, survival predictions depend on an accurate tumor burden. Deviations could misrepresent the prognosis and lead to incorrect treatment decisions.

Section 5 — Trust & Preference (Survival Prediction Focus)

5.1 How likely is it that AI-generated masks are more accurate than GT in some cases (for example, subtle lesions missed by physicians, or over-delineation that reduces noise) in survival prediction tasks?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = very unlikely, 5 = very likely)

Reason:

In some cases, AI-generated masks could be more accurate than GT, especially when over-delineation is used to reduce noise. However, the use of radiomic features by physicians could decrease this superiority.

5.2 Would you prefer to use the AI mask (rather than GT) as input for survival prediction?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not at all, 5 = strongly yes)

Reason:

I prefer GT masks for reasons explained in the above answers, but I may use AI masks in additional to GT masks.

Section 6 — Summaries and Overall Assessment

6.1 Diagnosis-oriented evaluation — How successful is the AI mask in lesion delineation compared to GT?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = complete failure, 5 = highly successful)

Reason:

AI masks are more useful for delineating larger lesions than smaller ones.

6.2 Prediction-oriented evaluation — How successful is the AI mask in contributing to survival prediction compared to GT?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no added value, 5 = very strong added value)

Reason:

The answer to this question requires multiple experiments and statistical analysis.

6.3 Overall confidence in your assessment

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no confidence, 5 = complete confidence)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6.4 Do you think geometric evaluation metrics (e.g., Dice, IoU, Hausdorff distance) are sufficient to establish clinical trust in AI segmentation?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not sufficient at all, 5 = fully sufficient)

Reason:

Geometric metrics alone are not sufficient to establish clinical trust in AI segmentation. Quantifying technical accuracy requires them, but they are not sufficient for measuring clinical relevance or reliability.

6.5 What additional types of evaluation do you think are necessary for trustworthy AI segmentation (e.g., radiomics stability, survival prediction, physician review, workflow impact, and others), and why do you think those are effective in building trustworthiness?

Reason:

Physicians should control AI masks because clinical assessment matters more than geometric metrics alone.

Section 7 — Clinical Adoption & Workflow Impact

7.1 Usability — How easy is it to interpret and use the AI mask compared to GT in clinical workflow?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = much harder, 5 = much easier)

Reason:

AI masks improve the usability of workflows that require speed, consistency, and high throughput, especially when physicians only use manual segmentation rather than Radiomic features.

7.2 Clinical actionability — Would the AI–GT differences change clinical decisions (for example, radiotherapy target, surgery extent, follow-up planning)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no impact, 5 = strong impact)

Reason:

Yes, differences between AI and GT masks can potentially impact clinical decisions. However, the impact depends on the location of the lesion, the type of lesion, and the clinical context.

7.3 Robustness and variability — How consistent and reliable are AI masks across different cases and observers?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = very inconsistent, 5 = very consistent)

Reason:

AI masks reduce interobserver variability, resulting in greater consistency across cases and observers. This is especially important when the observers are inexperienced.

7.4 Time-saving potential — How much time could this AI save compared to manual delineation?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no time saved, 5 = saves most effort)

Reason:

AI segmentation can significantly reduce delineation time, especially for complex cases involving multiple lesions.

7.5 Trust in clinical use — How confident would you be using AI masks for high-stakes tasks (e.g., radiotherapy planning) without physician refinement?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no trust, 5 = complete trust)

Reason:

For high-stakes tasks like radiotherapy planning, I would not fully trust AI masks without physician review because overestimation and underestimation by AI can cause damage to health tissue and insufficient treatment, respectively. Also, I miss small lesions that could alter the stage of the disease, which could in turn cause a change in treatment planning.

7.6 Confidence maps/interpretability (if available) — Do uncertainty or confidence maps improve your trust in AI outputs?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no improvement, 5 = very strong improvement)

Reason:

Yes, uncertainty or confidence maps can substantially improve trust in AI outputs, especially in clinical settings. These maps increase trust by making AI outputs transparent and risk-aware.

7.7 Role of AI in clinical practice — In your opinion, AI-based segmentation should:

☐ Replace physicians entirely  
☐ Provide an initial mask for physicians to refine  
☐ Assist physicians as a support/check tool

Reason:

This is because clinicians supervise high-stakes decisions.

**End**

**Comparative Clinical Assessment of Deep Learning Tumor Masks vs. Ground Truth in Lung Cancer**

Model under review: 3DMedSam

**Segmentation Results**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Analysis Mode | Model | Dice Score | IoU Score | Hausdorff Distance |
| Test |  |  |  |  |
|  |  |  |  |
|  |  |  |  |
| **3DMedSam** | 0.68 0.19 | 0.52 0.18 | 4.26 1.55 |

**Prediction Result:**

Radiomics extracted from Ground Truth used for prediction tasks

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Validation | | | | | | Test | | | | | | | Names | | | |
| Acc | Precision | Recall | F1 | AUC | Specificity | Acc | Precision | Recall | F1 | AUC | Specificity | FSA/AEA | | CA | Model | SL/SSL |
| 0.88 | 0.78 | 0.88 | 0.83 | 0.54 | 1 | 0.87 | 0.76 | 0.87 | 0.81 | 0.71 | 1 | Mutual\_Info\_Gain\_Ratio | | ExtraTree | GT | SSL |
| 0.74 | 0.67 | 0.74 | 0.66 | 0.55 | 0.98 | 0.87 | 0.79 | 0.87 | 0.83 | 0.63 | 0.99 | SL |

Radiomics extracted from DL 3DmedSam used for prediction tasks

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Validation | | | | | | Test | | | | | | Names | | | |
| Acc | Precision | Recall | F1 | AUC | Specificity | Acc | Precision | Recall | F1 | AUC | Specificity | FSA/AEA | CA | Model | SL/SSL |
| 0.88 | 0.77 | 0.88 | 0.82 | 0.67 | 1 | 0.85 | 0.71 | 0.85 | 0.77 | 0.68 | 1 | ReliefF | GradientBoosting | MedSam | SSL |
| 0.74 | 0.55 | 0.74 | 0.63 | 0.51 | 0.99 | 0.86 | 0.76 | 0.86 | 0.81 | 0.71 | 0.99 | SL |

**Start**

Section 1 — Overall Impression

1.1 Does the AI mask define a tumor volume that is more clinically meaningful than the Ground Truth (GT) for diagnosis and visual interpretation?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not at all, 5 = very strongly yes)

The AI mask does not provide a clinically more meaningful tumor volume definition than the ground truth (GT) for diagnosis and visual interpretation, especially in the necrotic regions of tumors.

Section 2 — Spatial Differences

2.1 Are the AI–GT differences concentrated in the peritumoral zone (≈5–15 mm around the tumor)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not at all, 5 = strongly concentrated)

Reason:

Additional data, such as information on image slice thickness, is needed to answer this question definitively. In general, however, the answer is yes, however, in this AI model, though, this comes at the cost of underestimating tumor volume in the periphery.

2.2 How clinically meaningful are the regions where the AI mask differs from GT (including over-segmentation or under-segmentation, such as peritumoral expansion, missed necrotic cores, spiculated margins)? Or, how clinically justified are the AI–GT differences (whether due to additional regions or excluded regions)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

Reason:

This difference is clinically significant, with a greater impact observed in the necrotic areas compared to peritumoral and spiculated margins.

Section 3 — Clinical Interpretation of Differences

3.1 How well does the AI mask preserve internal tumor heterogeneity (for example: solid vs. ground-glass opacity, necrotic vs. viable regions)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

Reason:

AI masks fail to preserve internal tumor heterogeneity, particularly in necrotic regions. This leads results in an overestimation of tumor volume in these areas.

3.2 Boundary quality: compared to GT, how noisy or fragmented are AI boundaries?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = much less noisy, 5 = much more noisy)

Boundary is less noisy in AI masks compared to GT masks.

Section 4 — Link to Survival Prediction

4.1 Do the AI–GT differences likely capture prognostically relevant tissue that improves survival prediction?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

Reason:

The answer to this question requires multiple experiments and statistical analysis.

4.2 How would you rate the effect of the AI’s deviation from GT (larger or smaller volume) on survival prediction?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = strongly harmful, 5 = strongly beneficial)

Reason:

In this AI model overestimation of tumor volume in the necrotic regions and underestimation of tumor volume in peripheral regions affect on survival prediction.

Section 5 — Trust & Preference (Survival Prediction Focus)

5.1 How likely is it that AI-generated masks are more accurate than GT in some cases (for example, subtle lesions missed by physicians, or over-delineation that reduces noise) in survival prediction tasks?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = very unlikely, 5 = very likely)

Reason:

In some cases, AI-generated masks could be more accurate than GT, especially when over-delineation is used to reduce noise. However, the use of radiomic features by physicians could decrease this superiority.

5.2 Would you prefer to use the AI mask (rather than GT) as input for survival prediction?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not at all, 5 = strongly yes)

Reason:

I prefer GT masks for reasons explained in the above answers, but I may use AI masks in additional to GT masks.

Section 6 — Summaries and Overall Assessment

6.1 Diagnosis-oriented evaluation — How successful is the AI mask in lesion delineation compared to GT?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = complete failure, 5 = highly successful)

Reason:

In this AI mask model, underestimation of tumor volume is depicted in the peripheral regions.

6.2 Prediction-oriented evaluation — How successful is the AI mask in contributing to survival prediction compared to GT?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no added value, 5 = very strong added value)

Reason:

The answer to this question requires multiple experiments and statistical analysis.

6.3 Overall confidence in your assessment

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no confidence, 5 = complete confidence)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6.4 Do you think geometric evaluation metrics (e.g., Dice, IoU, Hausdorff distance) are sufficient to establish clinical trust in AI segmentation?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not sufficient at all, 5 = fully sufficient)

Reason:

Geometric metrics alone are not sufficient to establish clinical trust in AI segmentation. Quantifying technical accuracy requires them, but they are not sufficient for measuring clinical relevance or reliability.

6.5 What additional types of evaluation do you think are necessary for trustworthy AI segmentation (e.g., radiomics stability, survival prediction, physician review, workflow impact, and others), and why do you think those are effective in building trustworthiness?

Reason:

Physicians should control AI masks because clinical assessment matters more than geometric metrics alone. Additionally, for spiculated and necrotic regions, radiomic features can be used to supplement the AI model.

Section 7 — Clinical Adoption & Workflow Impact

7.1 Usability — How easy is it to interpret and use the AI mask compared to GT in clinical workflow?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = much harder, 5 = much easier)

Reason:

AI masks improve the usability of workflows that require speed, consistency, and high throughput, especially when physicians only use manual segmentation rather than Radiomic features.

7.2 Clinical actionability — Would the AI–GT differences change clinical decisions (for example, radiotherapy target, surgery extent, follow-up planning)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no impact, 5 = strong impact)

Reason:

Yes, differences between AI and GT masks can potentially impact clinical decisions. However, the impact depends on the location of the lesion, the type of lesion, and the clinical context.

7.3 Robustness and variability — How consistent and reliable are AI masks across different cases and observers?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = very inconsistent, 5 = very consistent)

Reason:

AI masks reduce interobserver variability, resulting in greater consistency across cases and observers. This is especially important when the observers are inexperienced.

7.4 Time-saving potential — How much time could this AI save compared to manual delineation?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no time saved, 5 = saves most effort)

Reason:

AI segmentation can significantly reduce delineation time, especially for complex cases involving multiple lesions.

7.5 Trust in clinical use — How confident would you be using AI masks for high-stakes tasks (e.g., radiotherapy planning) without physician refinement?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no trust, 5 = complete trust)

Reason:

For high-stakes tasks like radiotherapy planning, I would not fully trust AI masks without physician review because overestimation and underestimation by AI can cause damage to health tissue and insufficient treatment, respectively. Also, I miss small lesions that could alter the stage of the disease, which could in turn cause a change in treatment planning.

7.6 Confidence maps/interpretability (if available) — Do uncertainty or confidence maps improve your trust in AI outputs?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no improvement, 5 = very strong improvement)

Reason:

Yes, uncertainty or confidence maps can substantially improve trust in AI outputs, especially in clinical settings. These maps increase trust by making AI outputs transparent and risk-aware.

7.7 Role of AI in clinical practice — In your opinion, AI-based segmentation should:

☐ Replace physicians entirely  
☐ Provide an initial mask for physicians to refine  
☐ Assist physicians as a support/check tool

Reason:

This is because clinicians supervise high-stakes decisions.

**End**

**Comparative Clinical Assessment of Deep Learning Tumor Masks vs. Ground Truth in Lung Cancer**

Model under review: Reconnet

**Segmentation Results**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Analysis Mode | Model | Dice Score | IoU Score | Hausdorff Distance |
| Test |  |  |  |  |
|  |  |  |  |
| **Reconnet** | 0.80 0.09 | 0.68 0.13 | 11.30 3.96 |
|  |  |  |  |
|  |  |  |  |

**Prediction Result:**

Radiomics extracted from Ground Truth used for prediction tasks

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Validation | | | | | | Test | | | | | | | Names | | | |
| Acc | Precision | Recall | F1 | AUC | Specificity | Acc | Precision | Recall | F1 | AUC | Specificity | FSA/AEA | | CA | Model | SL/SSL |
| 0.88 | 0.78 | 0.88 | 0.83 | 0.54 | 1 | 0.87 | 0.76 | 0.87 | 0.81 | 0.71 | 1 | Mutual\_Info\_Gain\_Ratio | | ExtraTree | GT | SSL |
| 0.74 | 0.67 | 0.74 | 0.66 | 0.55 | 0.98 | 0.87 | 0.79 | 0.87 | 0.83 | 0.63 | 0.99 | SL |

Radiomics extracted from DL Reconnet used for prediction tasks

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Validation | | | | | | Test | | | | | | Names | | | |
| Acc | Precision | Recall | F1 | AUC | Specificity | Acc | Precision | Recall | F1 | AUC | Specificity | FSA/AEA | CA | Model | SL/SSL |
| 0.88 | 0.78 | 0.88 | 0.83 | 0.65 | 1 | 0.86 | 0.75 | 0.86 | 0.80 | 0.71 | 1 | UMAP | AdaBoost | Reconnet | SSL |
| 0.75 | 0.56 | 0.75 | 0.64 | 0.50 | 1 | 0.88 | 0.77 | 0.88 | 0.82 | 0.5 | 1 | SL |

**Start**

Section 1 — Overall Impression

1.1 Does the AI mask define a tumor volume that is more clinically meaningful than the Ground Truth (GT) for diagnosis and visual interpretation?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not at all, 5 = very strongly yes)

The AI mask does not provide a clinically more meaningful tumor volume definition than the ground truth (GT) for diagnosis and visual interpretation, especially in the necrotic regions of tumors. Misregistration is also evident in certain cases.

Section 2 — Spatial Differences

2.1 Are the AI–GT differences concentrated in the peritumoral zone (≈5–15 mm around the tumor)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not at all, 5 = strongly concentrated)

Reason:

More data is needed to answer this question correctly, such as information about the size of the image slices. In general, however, the answer is no, especially for spiculated lesions. Additionally, in some cases, underestimation is evident in certain peripheral tumor regions.

2.2 How clinically meaningful are the regions where the AI mask differs from GT (including over-segmentation or under-segmentation, such as peritumoral expansion, missed necrotic cores, spiculated margins)? Or, how clinically justified are the AI–GT differences (whether due to additional regions or excluded regions)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

Reason:

This difference is clinically significant, with a greater impact observed in the necrotic, peritumoral and spiculated margins.

Section 3 — Clinical Interpretation of Differences

3.1 How well does the AI mask preserve internal tumor heterogeneity (for example: solid vs. ground-glass opacity, necrotic vs. viable regions)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

Reason:

AI-generated masks often fail to preserve internal tumor heterogeneity, particularly in necrotic regions, leading to an overestimation of tumor volume in these areas. In contrast, underestimation may occur at the peripheral margins in some cases.

3.2 Boundary quality: compared to GT, how noisy or fragmented are AI boundaries?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = much less noisy, 5 = much more noisy)

Boundary is less noisy in AI masks compared to GT masks.

Section 4 — Link to Survival Prediction

4.1 Do the AI–GT differences likely capture prognostically relevant tissue that improves survival prediction?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

Reason:

The answer to this question requires multiple experiments and statistical analysis.

4.2 How would you rate the effect of the AI’s deviation from GT (larger or smaller volume) on survival prediction?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = strongly harmful, 5 = strongly beneficial)

Reason:

In this AI model overestimation of tumor volume in the necrotic regions and underestimation of tumor volume in peripheral regions affect on survival prediction.

Section 5 — Trust & Preference (Survival Prediction Focus)

5.1 How likely is it that AI-generated masks are more accurate than GT in some cases (for example, subtle lesions missed by physicians, or over-delineation that reduces noise) in survival prediction tasks?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = very unlikely, 5 = very likely)

Reason:

In some cases, AI-generated masks could be more accurate than GT, especially when over-delineation is used to reduce noise. However, the use of radiomic features by physicians could decrease this superiority.

5.2 Would you prefer to use the AI mask (rather than GT) as input for survival prediction?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not at all, 5 = strongly yes)

Reason:

I prefer GT masks for reasons explained in the above answers, but I may use AI masks in additional to GT masks.

Section 6 — Summaries and Overall Assessment

6.1 Diagnosis-oriented evaluation — How successful is the AI mask in lesion delineation compared to GT?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = complete failure, 5 = highly successful)

Reason:

In this AI mask model, underestimation of tumor volume is depicted in the peripheral regions.

6.2 Prediction-oriented evaluation — How successful is the AI mask in contributing to survival prediction compared to GT?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no added value, 5 = very strong added value)

Reason:

The answer to this question requires multiple experiments and statistical analysis.

6.3 Overall confidence in your assessment

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no confidence, 5 = complete confidence)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6.4 Do you think geometric evaluation metrics (e.g., Dice, IoU, Hausdorff distance) are sufficient to establish clinical trust in AI segmentation?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not sufficient at all, 5 = fully sufficient)

Reason:

Geometric metrics alone are not sufficient to establish clinical trust in AI segmentation. Quantifying technical accuracy requires them, but they are not sufficient for measuring clinical relevance or reliability.

6.5 What additional types of evaluation do you think are necessary for trustworthy AI segmentation (e.g., radiomics stability, survival prediction, physician review, workflow impact, and others), and why do you think those are effective in building trustworthiness?

Reason:

Physicians should control AI masks because clinical assessment matters more than geometric metrics alone. Additionally, for spiculated and necrotic regions, radiomic features can be used to supplement the AI model.

Section 7 — Clinical Adoption & Workflow Impact

7.1 Usability — How easy is it to interpret and use the AI mask compared to GT in clinical workflow?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = much harder, 5 = much easier)

Reason:

AI masks improve the usability of workflows that require speed, consistency, and high throughput, especially when physicians only use manual segmentation rather than Radiomic features.

7.2 Clinical actionability — Would the AI–GT differences change clinical decisions (for example, radiotherapy target, surgery extent, follow-up planning)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no impact, 5 = strong impact)

Reason:

Yes, differences between AI and GT masks can potentially impact clinical decisions. However, the impact depends on the location of the lesion, the type of lesion, and the clinical context.

7.3 Robustness and variability — How consistent and reliable are AI masks across different cases and observers?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = very inconsistent, 5 = very consistent)

Reason:

AI masks reduce interobserver variability, resulting in greater consistency across cases and observers. This is especially important when the observers are inexperienced.

7.4 Time-saving potential — How much time could this AI save compared to manual delineation?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no time saved, 5 = saves most effort)

Reason:

AI segmentation can significantly reduce delineation time, especially for complex cases involving multiple lesions.

7.5 Trust in clinical use — How confident would you be using AI masks for high-stakes tasks (e.g., radiotherapy planning) without physician refinement?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no trust, 5 = complete trust)

Reason:

For high-stakes tasks like radiotherapy planning, I would not fully trust AI masks without physician review because overestimation and underestimation by AI can cause damage to health tissue and insufficient treatment, respectively. Additionally, I might miss small lesions that could change the disease stage, potentially altering the treatment plan.

7.6 Confidence maps/interpretability (if available) — Do uncertainty or confidence maps improve your trust in AI outputs?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no improvement, 5 = very strong improvement)

Reason:

Yes, uncertainty or confidence maps can substantially improve trust in AI outputs, especially in clinical settings. These maps increase trust by making AI outputs transparent and risk-aware.

7.7 Role of AI in clinical practice — In your opinion, AI-based segmentation should:

☐ Replace physicians entirely  
☐ Provide an initial mask for physicians to refine  
☐ Assist physicians as a support/check tool

Reason:

This is because clinicians supervise high-stakes decisions.

**End**

**Comparative Clinical Assessment of Deep Learning Tumor Masks vs. Ground Truth in Lung Cancer**

Model under review: ResUnet

**Segmentation Results**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Analysis Mode | Model | Dice Score | IoU Score | Hausdorff Distance |
| Test |  |  |  |  |
| **ResUNet** | 75 0.08 | 0.61 0.10 | 17.95 6.91 |
|  |  |  |  |

**Prediction Result:**

Radiomics extracted from Ground Truth used for prediction tasks

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Validation | | | | | | Test | | | | | | | Names | | | |
| Acc | Precision | Recall | F1 | AUC | Specificity | Acc | Precision | Recall | F1 | AUC | Specificity | FSA/AEA | | CA | Model | SL/SSL |
| 0.88 | 0.78 | 0.88 | 0.83 | 0.54 | 1 | 0.87 | 0.76 | 0.87 | 0.81 | 0.71 | 1 | Mutual\_Info\_Gain\_Ratio | | ExtraTree | GT | SSL |
| 0.74 | 0.67 | 0.74 | 0.66 | 0.55 | 0.98 | 0.87 | 0.79 | 0.87 | 0.83 | 0.63 | 0.99 | SL |

Radiomics extracted from DL ResUNet used for prediction tasks

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Validation | | | | | | Test | | | | | | | Names | | | |
| Acc | Precision | Recall | F1 | AUC | Specificity | Acc | Precision | Recall | F1 | AUC | Specificity | FSA/AEA | | CA | Model | SL/SSL |
| 0.88 | 0.78 | 0.88 | 0.83 | 0.30 | 1 | 0.84 | 0.71 | 0.84 | 0.77 | 0.67 | 1 | Mutual Information | | MLP | ResUNet | SSL |
| 0.75 | 0.56 | 0.75 | 0.64 | 0.56 | 1 | 0.83 | 0.69 | 0.83 | 0.75 | 0.68 | 1 | SL |

**Start**

Section 1 — Overall Impression

1.1 Does the AI mask define a tumor volume that is more clinically meaningful than the Ground Truth (GT) for diagnosis and visual interpretation?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not at all, 5 = very strongly yes)

The AI-generated mask does not offer a clinically more meaningful definition of tumor volume compared to the ground truth (GT) for diagnostic or visual interpretation, particularly in the spiculated regions of tumors. Misregistration is also apparent in some cases.

Section 2 — Spatial Differences

2.1 Are the AI–GT differences concentrated in the peritumoral zone (≈5–15 mm around the tumor)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not at all, 5 = strongly concentrated)

Reason:

More data are needed to answer this question accurately, such as information about the size of the image slices. In general, however, the answer is no—especially for spiculated lesions and craniocaudal borders of lesions.

2.2 How clinically meaningful are the regions where the AI mask differs from GT (including over-segmentation or under-segmentation, such as peritumoral expansion, missed necrotic cores, spiculated margins)? Or, how clinically justified are the AI–GT differences (whether due to additional regions or excluded regions)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

Reason:

This difference is clinically significant, with the greatest impact observed in the craniocaudal and spiculated margins.

Section 3 — Clinical Interpretation of Differences

3.1 How well does the AI mask preserve internal tumor heterogeneity (for example: solid vs. ground-glass opacity, necrotic vs. viable regions)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

Reason:

AI-generated masks often fail to preserve internal tumor heterogeneity, particularly in necrotic regions. The differentiation of heterogeneity depends on necrotic region size, with AI models showing less heterogeneity differentiation in smaller necrotic lesions.

3.2 Boundary quality: compared to GT, how noisy or fragmented are AI boundaries?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = much less noisy, 5 = much more noisy)

Boundary is less noisy in AI masks compared to GT masks.

Section 4 — Link to Survival Prediction

4.1 Do the AI–GT differences likely capture prognostically relevant tissue that improves survival prediction?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

Reason:

The answer to this question requires multiple experiments and statistical analysis.

4.2 How would you rate the effect of the AI’s deviation from GT (larger or smaller volume) on survival prediction?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = strongly harmful, 5 = strongly beneficial)

Reason:

In this AI model overestimation of tumor volume in the necrotic regions and underestimation of tumor volume in peripheral regions affect on survival prediction.

Section 5 — Trust & Preference (Survival Prediction Focus)

5.1 How likely is it that AI-generated masks are more accurate than GT in some cases (for example, subtle lesions missed by physicians, or over-delineation that reduces noise) in survival prediction tasks?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = very unlikely, 5 = very likely)

Reason:

In some cases, AI-generated masks could be more accurate than GT, especially when over-delineation is used to reduce noise. However, the use of radiomic features by physicians could decrease this superiority.

5.2 Would you prefer to use the AI mask (rather than GT) as input for survival prediction?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not at all, 5 = strongly yes)

Reason:

I prefer GT masks for reasons explained in the above answers, but I may use AI masks in additional to GT masks.

Section 6 — Summaries and Overall Assessment

6.1 Diagnosis-oriented evaluation — How successful is the AI mask in lesion delineation compared to GT?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = complete failure, 5 = highly successful)

Reason:

In this AI mask model, misregistration of tumor volume is observed at the craniocaudal borders.

6.2 Prediction-oriented evaluation — How successful is the AI mask in contributing to survival prediction compared to GT?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no added value, 5 = very strong added value)

Reason:

The answer to this question requires multiple experiments and statistical analysis.

6.3 Overall confidence in your assessment

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no confidence, 5 = complete confidence)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6.4 Do you think geometric evaluation metrics (e.g., Dice, IoU, Hausdorff distance) are sufficient to establish clinical trust in AI segmentation?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not sufficient at all, 5 = fully sufficient)

Reason:

Geometric metrics alone are not sufficient to establish clinical trust in AI segmentation. Quantifying technical accuracy requires them, but they are not sufficient for measuring clinical relevance or reliability.

6.5 What additional types of evaluation do you think are necessary for trustworthy AI segmentation (e.g., radiomics stability, survival prediction, physician review, workflow impact, and others), and why do you think those are effective in building trustworthiness?

Reason:

Physicians should control AI masks because clinical assessment matters more than geometric metrics alone. Additionally, for spiculated and necrotic regions, radiomic features can be used to supplement the AI model.

Section 7 — Clinical Adoption & Workflow Impact

7.1 Usability — How easy is it to interpret and use the AI mask compared to GT in clinical workflow?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = much harder, 5 = much easier)

Reason:

AI masks improve the usability of workflows that require speed, consistency, and high throughput, especially when physicians only use manual segmentation rather than Radiomic features.

7.2 Clinical actionability — Would the AI–GT differences change clinical decisions (for example, radiotherapy target, surgery extent, follow-up planning)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no impact, 5 = strong impact)

Reason:

Yes, differences between AI and GT masks can potentially impact clinical decisions. However, the impact depends on the location of the lesion, the type of lesion, and the clinical context.

7.3 Robustness and variability — How consistent and reliable are AI masks across different cases and observers?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = very inconsistent, 5 = very consistent)

Reason:

AI masks reduce interobserver variability, resulting in greater consistency across cases and observers. This is especially important when the observers are inexperienced.

7.4 Time-saving potential — How much time could this AI save compared to manual delineation?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no time saved, 5 = saves most effort)

Reason:

AI segmentation can significantly reduce delineation time, especially for complex cases involving multiple lesions.

7.5 Trust in clinical use — How confident would you be using AI masks for high-stakes tasks (e.g., radiotherapy planning) without physician refinement?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no trust, 5 = complete trust)

Reason:

For high-stakes tasks like radiotherapy planning, I would not fully trust AI masks without physician review because overestimation and underestimation by AI can cause damage to health tissue and insufficient treatment, respectively. Additionally, I might miss small lesions that could change the disease stage, potentially altering the treatment plan.

7.6 Confidence maps/interpretability (if available) — Do uncertainty or confidence maps improve your trust in AI outputs?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no improvement, 5 = very strong improvement)

Reason:

Yes, uncertainty or confidence maps can substantially improve trust in AI outputs, especially in clinical settings. These maps increase trust by making AI outputs transparent and risk-aware.

7.7 Role of AI in clinical practice — In your opinion, AI-based segmentation should:

☐ Replace physicians entirely  
☐ Provide an initial mask for physicians to refine  
☐ Assist physicians as a support/check tool

Reason:

This is because clinicians supervise high-stakes decisions.

**End**

**Comparative Clinical Assessment of Deep Learning Tumor Masks vs. Ground Truth in Lung Cancer**

Model under review: VNet

**Segmentation Results**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Analysis Mode | Model | Dice Score | IoU Score | Hausdorff Distance |
| Test | **VNet** | 0.83 0.07 | 0.71 0.09 | 10.04 3.73 |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

**Prediction Result:**

Radiomics extracted from Ground Truth used for prediction tasks

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Validation | | | | | | Test | | | | | | | Names | | | |
| Acc | Precision | Recall | F1 | AUC | Specificity | Acc | Precision | Recall | F1 | AUC | Specificity | FSA/AEA | | CA | Model | SL/SSL |
| 0.88 | 0.78 | 0.88 | 0.83 | 0.54 | 1 | 0.87 | 0.76 | 0.87 | 0.81 | 0.71 | 1 | Mutual\_Info\_Gain\_Ratio | | ExtraTree | GT | SSL |
| 0.74 | 0.67 | 0.74 | 0.66 | 0.55 | 0.98 | 0.87 | 0.79 | 0.87 | 0.83 | 0.63 | 0.99 | SL |

Radiomics extracted from Vnet used for prediction tasks

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Validation | | | | | | Test | | | | | | Names | | | |
| Acc | Precision | Recall | F1 | AUC | Specificity | Acc | Precision | Recall | F1 | AUC | Specificity | FSA/AEA | CA | Model | SL/SSL |
| 0.88 | 0.78 | 0.88 | 0.83 | 0.65 | 1 | 0.88 | 0.77 | 0.88 | 0.82 | 0.76 | 1 | LASSO | LGBM | VNet | SSL |
| 0.74 | 0.55 | 0.74 | 0.63 | 0.58 | 0.99 | 0.87 | 0.77 | 0.87 | 0.81 | 0.40 | 0.99 | SL |

**Start**

Section 1 — Overall Impression

1.1 Does the AI mask define a tumor volume that is more clinically meaningful than the Ground Truth (GT) for diagnosis and visual interpretation?

0 ☐ 1 ☐ 2 **☐** 3 ☐ 4 ☐ 5 ☐ (0 = not at all, 5 = very strongly yes)

Reason:

The AI mask does not provide a clinically more meaningful tumor volume definition than the ground truth (GT) for diagnosis and visual interpretation, especially in the case of irregular, spiculated borders and necrotic regions of tumors.

Section 2 — Spatial Differences

2.1 Are the AI–GT differences concentrated in the peritumoral zone (≈5–15 mm around the tumor)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not at all, 5 = strongly concentrated)

Reason:

More data is needed to answer this question correctly, such as the size of the image slices. In general, however, the answer is no, especially for spiculated lesions.

2.2 How clinically meaningful are the regions where the AI mask differs from GT (including over-segmentation or under-segmentation, such as peritumoral expansion, missed necrotic cores, spiculated margins)? Or, how clinically justified are the AI–GT differences (whether due to additional regions or excluded regions)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

Reason:

This difference is more clinically significant in necrotic areas than in areas with peritumoral and spiculated margins. However, peritumoral extension near the chest wall is also clinically significant because it changes the stage of the disease.

Section 3 — Clinical Interpretation of Differences

3.1 How well does the AI mask preserve internal tumor heterogeneity (for example: solid vs. ground-glass opacity, necrotic vs. viable regions)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

Reason:

AI masks fail to preserve internal tumor heterogeneity, resulting in an overestimation of tumor volume in ground-glass opacity and necrotic regions.

3.2 Boundary quality: compared to GT, how noisy or fragmented are AI boundaries?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = much less noisy, 5 = much more noisy)

Reason:

No significant difference in boundary quality is observed between AI and GT.

Section 4 — Link to Survival Prediction

4.1 Do the AI–GT differences likely capture prognostically relevant tissue that improves survival prediction?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

Reason:

The answer to this question requires multiple experiments and statistical analysis.

4.2 How would you rate the effect of the AI’s deviation from GT (larger or smaller volume) on survival prediction?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = strongly harmful, 5 = strongly beneficial)

Reason:

Depending on the tumor's location and the type of lesions, survival predictions depend on an accurate tumor burden. Deviations could misrepresent the prognosis and lead to incorrect treatment decisions.

Section 5 — Trust & Preference (Survival Prediction Focus)

5.1 How likely is it that AI-generated masks are more accurate than GT in some cases (for example, subtle lesions missed by physicians, or over-delineation that reduces noise) in survival prediction tasks?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = very unlikely, 5 = very likely)

Reason:

In some cases, AI-generated masks could be more accurate than GT, especially when over-delineation is used to reduce noise. However, the use of radiomic features by physicians could decrease this superiority.

5.2 Would you prefer to use the AI mask (rather than GT) as input for survival prediction?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not at all, 5 = strongly yes)

Reason:

I prefer GT masks for reasons explained in the above answers, but I may use AI masks in additional to GT masks.

Section 6 — Summaries and Overall Assessment

6.1 Diagnosis-oriented evaluation — How successful is the AI mask in lesion delineation compared to GT?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = complete failure, 5 = highly successful)

Reason:

AI masks are more useful for delineating larger lesions than smaller ones.

6.2 Prediction-oriented evaluation — How successful is the AI mask in contributing to survival prediction compared to GT?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no added value, 5 = very strong added value)

Reason:

The answer to this question requires multiple experiments and statistical analysis.

6.3 Overall confidence in your assessment

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no confidence, 5 = complete confidence)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6.4 Do you think geometric evaluation metrics (e.g., Dice, IoU, Hausdorff distance) are sufficient to establish clinical trust in AI segmentation?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not sufficient at all, 5 = fully sufficient)

Reason:

Geometric metrics alone are not sufficient to establish clinical trust in AI segmentation. Quantifying technical accuracy requires them, but they are not sufficient for measuring clinical relevance or reliability.

6.5 What additional types of evaluation do you think are necessary for trustworthy AI segmentation (e.g., radiomics stability, survival prediction, physician review, workflow impact, and others), and why do you think those are effective in building trustworthiness?

Reason:

Physicians should control AI masks because clinical assessment matters more than geometric metrics alone.

Section 7 — Clinical Adoption & Workflow Impact

7.1 Usability — How easy is it to interpret and use the AI mask compared to GT in clinical workflow?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = much harder, 5 = much easier)

Reason:

AI masks improve the usability of workflows that require speed, consistency, and high throughput, especially when physicians only use manual segmentation rather than Radiomic features.

7.2 Clinical actionability — Would the AI–GT differences change clinical decisions (for example, radiotherapy target, surgery extent, follow-up planning)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no impact, 5 = strong impact)

Reason:

Yes, differences between AI and GT masks can potentially impact clinical decisions. However, the impact depends on the location of the lesion, the type of lesion, and the clinical context.

7.3 Robustness and variability — How consistent and reliable are AI masks across different cases and observers?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = very inconsistent, 5 = very consistent)

Reason:

AI masks reduce human-induced variability, resulting in greater consistency across cases and observers, which is particularly important when observers have less experience.

7.4 Time-saving potential — How much time could this AI save compared to manual delineation?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no time saved, 5 = saves most effort)

Reason:

AI segmentation can significantly reduce the time spent on delineation, particularly for complex cases involving multiple lesions.

7.5 Trust in clinical use — How confident would you be using AI masks for high-stakes tasks (e.g., radiotherapy planning) without physician refinement?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no trust, 5 = complete trust)

Reason:

For high-stakes tasks like radiotherapy planning, I would not fully trust AI masks without physician review because overestimation and underestimation by AI can cause damage to health tissue and insufficient treatment, respectively. Also, I miss small lesions that could alter the stage of the disease, which could in turn cause a change in treatment planning.

7.6 Confidence maps/interpretability (if available) — Do uncertainty or confidence maps improve your trust in AI outputs?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no improvement, 5 = very strong improvement)

Reason:

Yes, uncertainty or confidence maps can substantially improve trust in AI outputs, especially in clinical settings. These maps increase trust by making AI outputs transparent and risk-aware.

7.7 Role of AI in clinical practice — In your opinion, AI-based segmentation should:

☐ Replace physicians entirely  
☐ Provide an initial mask for physicians to refine  
☐ Assist physicians as a support/check tool

Reason:

This is because clinicians supervise high-stakes decisions.

**End**

**D3**

**Comparative Clinical Assessment of Deep Learning Tumor Masks vs. Ground Truth in Lung Cancer**

Model under review: 3DAttention

**Segmentation Results**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Analysis Mode | Model | Dice Score | IoU Score | Hausdorff Distance |
| Test |  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
| **3DAttention** | 0.83 0.04 | 0.72 0.04 | 12.12 3.73 |

**Prediction Result:**

Radiomics extracted from Ground Truth used for prediction tasks

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Validation | | | | | | Test | | | | | | | Names | | | |
| Acc | Precision | Recall | F1 | AUC | Specificity | Acc | Precision | Recall | F1 | AUC | Specificity | FSA/AEA | | CA | Model | SL/SSL |
| 0.88 | 0.78 | 0.88 | 0.83 | 0.54 | 1 | 0.87 | 0.76 | 0.87 | 0.81 | 0.71 | 1 | Mutual\_Info\_Gain\_Ratio | | ExtraTree | GT | SSL |
| 0.74 | 0.67 | 0.74 | 0.66 | 0.55 | 0.98 | 0.87 | 0.79 | 0.87 | 0.83 | 0.63 | 0.99 | SL |

Radiomics extracted from DL 3D attention used for prediction tasks

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Validation | | | | | | Test | | | | | | | Names | | | |
| Acc | Precision | Recall | F1 | AUC | Specificity | Acc | Precision | Recall | F1 | AUC | Specificity | FSA/AEA | | CA | Model | SL/SSL |
| 0.79 | 0.80 | 0.79 | 0.80 | 0.69 | 0.87 | 0.81 | 0.83 | 0.81 | 0.82 | 0.76 | 0.88 | Guassian Random Projection | | Voting Classifier | 3DAtt | SSL |
| 0.61 | 0.64 | 0.61 | 0.62 | 0.53 | 0.69 | 0.72 | 0.79 | 0.72 | 0.75 | 0.55 | 0.79 | SL |

**Start**

Section 1 — Overall Impression

1.1 Does the AI mask define a tumor volume that is more clinically meaningful than the Ground Truth (GT) for diagnosis and visual interpretation?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not at all, 5 = very strongly yes)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 2 — Spatial Differences

2.1 Are the AI–GT differences concentrated in the peritumoral zone (≈5–15 mm around the tumor)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not at all, 5 = strongly concentrated)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

2.2 How clinically meaningful are the regions where the AI mask differs from GT (including over-segmentation or under-segmentation, such as peritumoral expansion, missed necrotic cores, spiculated margins)? Or, how clinically justified are the AI–GT differences (whether due to additional regions or excluded regions)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 3 — Clinical Interpretation of Differences

3.1 How well does the AI mask preserve internal tumor heterogeneity (for example: solid vs. ground-glass opacity, necrotic vs. viable regions)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

3.2 Boundary quality: compared to GT, how noisy or fragmented are AI boundaries?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = much less noisy, 5 = much more noisy)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 4 — Link to Survival Prediction

4.1 Do the AI–GT differences likely capture prognostically relevant tissue that improves survival prediction?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

4.2 How would you rate the effect of the AI’s deviation from GT (larger or smaller volume) on survival prediction?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = strongly harmful, 5 = strongly beneficial)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 5 — Trust & Preference (Survival Prediction Focus)

5.1 How likely is it that AI-generated masks are more accurate than GT in some cases (for example, subtle lesions missed by physicians, or over-delineation that reduces noise) in survival prediction tasks?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = very unlikely, 5 = very likely)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

5.2 Would you prefer to use the AI mask (rather than GT) as input for survival prediction?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not at all, 5 = strongly yes)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 6 — Summaries and Overall Assessment

6.1 Diagnosis-oriented evaluation — How successful is the AI mask in lesion delineation compared to GT?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = complete failure, 5 = highly successful)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6.2 Prediction-oriented evaluation — How successful is the AI mask in contributing to survival prediction compared to GT?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no added value, 5 = very strong added value)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6.3 Overall confidence in your assessment

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no confidence, 5 = complete confidence)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6.4 Do you think geometric evaluation metrics (e.g., Dice, IoU, Hausdorff distance) are sufficient to establish clinical trust in AI segmentation?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not sufficient at all, 5 = fully sufficient)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6.5 What additional types of evaluation do you think are necessary for trustworthy AI segmentation (e.g., radiomics stability, survival prediction, physician review, workflow impact, and others), and why do you think those are effective in building trustworthiness?

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 7 — Clinical Adoption & Workflow Impact

7.1 Usability — How easy is it to interpret and use the AI mask compared to GT in clinical workflow?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = much harder, 5 = much easier)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.2 Clinical actionability — Would the AI–GT differences change clinical decisions (for example, radiotherapy target, surgery extent, follow-up planning)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no impact, 5 = strong impact)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.3 Robustness and variability — How consistent and reliable are AI masks across different cases and observers?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = very inconsistent, 5 = very consistent)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.4 Time-saving potential — How much time could this AI save compared to manual delineation?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no time saved, 5 = saves most effort)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.5 Trust in clinical use — How confident would you be using AI masks for high-stakes tasks (e.g., radiotherapy planning) without physician refinement?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no trust, 5 = complete trust)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.6 Confidence maps/interpretability (if available) — Do uncertainty or confidence maps improve your trust in AI outputs?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no improvement, 5 = very strong improvement)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.7 Role of AI in clinical practice — In your opinion, AI-based segmentation should:

☐ Replace physicians entirely  
☐ Provide an initial mask for physicians to refine  
☐ Assist physicians as a support/check tool

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**End**

**Comparative Clinical Assessment of Deep Learning Tumor Masks vs. Ground Truth in Lung Cancer**

Model under review: 3DMedSam

**Segmentation Results**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Analysis Mode | Model | Dice Score | IoU Score | Hausdorff Distance |
| Test |  |  |  |  |
|  |  |  |  |
|  |  |  |  |
| **3DMedSam** | 0.68 0.19 | 0.52 0.18 | 4.26 1.55 |

**Prediction Result:**

Radiomics extracted from Ground Truth used for prediction tasks

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Validation | | | | | | Test | | | | | | | Names | | | |
| Acc | Precision | Recall | F1 | AUC | Specificity | Acc | Precision | Recall | F1 | AUC | Specificity | FSA/AEA | | CA | Model | SL/SSL |
| 0.88 | 0.78 | 0.88 | 0.83 | 0.54 | 1 | 0.87 | 0.76 | 0.87 | 0.81 | 0.71 | 1 | Mutual\_Info\_Gain\_Ratio | | ExtraTree | GT | SSL |
| 0.74 | 0.67 | 0.74 | 0.66 | 0.55 | 0.98 | 0.87 | 0.79 | 0.87 | 0.83 | 0.63 | 0.99 | SL |

Radiomics extracted from DL 3DmedSam used for prediction tasks

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Validation | | | | | | Test | | | | | | Names | | | |
| Acc | Precision | Recall | F1 | AUC | Specificity | Acc | Precision | Recall | F1 | AUC | Specificity | FSA/AEA | CA | Model | SL/SSL |
| 0.88 | 0.77 | 0.88 | 0.82 | 0.67 | 1 | 0.85 | 0.71 | 0.85 | 0.77 | 0.68 | 1 | ReliefF | GradientBoosting | MedSam | SSL |
| 0.74 | 0.55 | 0.74 | 0.63 | 0.51 | 0.99 | 0.86 | 0.76 | 0.86 | 0.81 | 0.71 | 0.99 | SL |

**Start**

Section 1 — Overall Impression

1.1 Does the AI mask define a tumor volume that is more clinically meaningful than the Ground Truth (GT) for diagnosis and visual interpretation?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not at all, 5 = very strongly yes)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 2 — Spatial Differences

2.1 Are the AI–GT differences concentrated in the peritumoral zone (≈5–15 mm around the tumor)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not at all, 5 = strongly concentrated)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

2.2 How clinically meaningful are the regions where the AI mask differs from GT (including over-segmentation or under-segmentation, such as peritumoral expansion, missed necrotic cores, spiculated margins)? Or, how clinically justified are the AI–GT differences (whether due to additional regions or excluded regions)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 3 — Clinical Interpretation of Differences

3.1 How well does the AI mask preserve internal tumor heterogeneity (for example: solid vs. ground-glass opacity, necrotic vs. viable regions)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

3.2 Boundary quality: compared to GT, how noisy or fragmented are AI boundaries?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = much less noisy, 5 = much more noisy)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 4 — Link to Survival Prediction

4.1 Do the AI–GT differences likely capture prognostically relevant tissue that improves survival prediction?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

4.2 How would you rate the effect of the AI’s deviation from GT (larger or smaller volume) on survival prediction?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = strongly harmful, 5 = strongly beneficial)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 5 — Trust & Preference (Survival Prediction Focus)

5.1 How likely is it that AI-generated masks are more accurate than GT in some cases (for example, subtle lesions missed by physicians, or over-delineation that reduces noise) in survival prediction tasks?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = very unlikely, 5 = very likely)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

5.2 Would you prefer to use the AI mask (rather than GT) as input for survival prediction?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not at all, 5 = strongly yes)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 6 — Summaries and Overall Assessment

6.1 Diagnosis-oriented evaluation — How successful is the AI mask in lesion delineation compared to GT?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = complete failure, 5 = highly successful)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6.2 Prediction-oriented evaluation — How successful is the AI mask in contributing to survival prediction compared to GT?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no added value, 5 = very strong added value)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6.3 Overall confidence in your assessment

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no confidence, 5 = complete confidence)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6.4 Do you think geometric evaluation metrics (e.g., Dice, IoU, Hausdorff distance) are sufficient to establish clinical trust in AI segmentation?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not sufficient at all, 5 = fully sufficient)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6.5 What additional types of evaluation do you think are necessary for trustworthy AI segmentation (e.g., radiomics stability, survival prediction, physician review, workflow impact, and others), and why do you think those are effective in building trustworthiness?

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 7 — Clinical Adoption & Workflow Impact

7.1 Usability — How easy is it to interpret and use the AI mask compared to GT in clinical workflow?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = much harder, 5 = much easier)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.2 Clinical actionability — Would the AI–GT differences change clinical decisions (for example, radiotherapy target, surgery extent, follow-up planning)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no impact, 5 = strong impact)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.3 Robustness and variability — How consistent and reliable are AI masks across different cases and observers?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = very inconsistent, 5 = very consistent)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.4 Time-saving potential — How much time could this AI save compared to manual delineation?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no time saved, 5 = saves most effort)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.5 Trust in clinical use — How confident would you be using AI masks for high-stakes tasks (e.g., radiotherapy planning) without physician refinement?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no trust, 5 = complete trust)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.6 Confidence maps/interpretability (if available) — Do uncertainty or confidence maps improve your trust in AI outputs?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no improvement, 5 = very strong improvement)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.7 Role of AI in clinical practice — In your opinion, AI-based segmentation should:

☐ Replace physicians entirely  
☐ Provide an initial mask for physicians to refine  
☐ Assist physicians as a support/check tool

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**End**

# **Comparative Clinical Assessment of Deep Learning Tumor Masks vs. Ground Truth in Lung Cancer**

Model under review: Reconnet

**Segmentation Results**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Analysis Mode | Model | Dice Score | IoU Score | Hausdorff Distance |
| Test |  |  |  |  |
|  |  |  |  |
| **Reconnet** | 0.80 0.09 | 0.68 0.13 | 11.30 3.96 |
|  |  |  |  |
|  |  |  |  |

**Prediction Result:**

Radiomics extracted from Ground Truth used for prediction tasks

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Validation | | | | | | Test | | | | | | | Names | | | |
| Acc | Precision | Recall | F1 | AUC | Specificity | Acc | Precision | Recall | F1 | AUC | Specificity | FSA/AEA | | CA | Model | SL/SSL |
| 0.88 | 0.78 | 0.88 | 0.83 | 0.54 | 1 | 0.87 | 0.76 | 0.87 | 0.81 | 0.71 | 1 | Mutual\_Info\_Gain\_Ratio | | ExtraTree | GT | SSL |
| 0.74 | 0.67 | 0.74 | 0.66 | 0.55 | 0.98 | 0.87 | 0.79 | 0.87 | 0.83 | 0.63 | 0.99 | SL |

Radiomics extracted from DL Reconnet used for prediction tasks

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Validation | | | | | | Test | | | | | | Names | | | |
| Acc | Precision | Recall | F1 | AUC | Specificity | Acc | Precision | Recall | F1 | AUC | Specificity | FSA/AEA | CA | Model | SL/SSL |
| 0.88 | 0.78 | 0.88 | 0.83 | 0.65 | 1 | 0.86 | 0.75 | 0.86 | 0.80 | 0.71 | 1 | UMAP | AdaBoost | Reconnet | SSL |
| 0.75 | 0.56 | 0.75 | 0.64 | 0.50 | 1 | 0.88 | 0.77 | 0.88 | 0.82 | 0.5 | 1 | SL |

**Start**

## Section 1 — Overall Impression

1.1 Does the AI mask define a tumor volume that is more clinically meaningful than the Ground Truth (GT) for diagnosis and visual interpretation?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not at all, 5 = very strongly yes)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

## Section 2 — Spatial Differences

2.1 Are the AI–GT differences concentrated in the peritumoral zone (≈5–15 mm around the tumor)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not at all, 5 = strongly concentrated)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

2.2 How clinically meaningful are the regions where the AI mask differs from GT (including over-segmentation or under-segmentation, such as peritumoral expansion, missed necrotic cores, spiculated margins)? Or, how clinically justified are the AI–GT differences (whether due to additional regions or excluded regions)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

## Section 3 — Clinical Interpretation of Differences

3.1 How well does the AI mask preserve internal tumor heterogeneity (for example: solid vs. ground-glass opacity, necrotic vs. viable regions)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

3.2 Boundary quality: compared to GT, how noisy or fragmented are AI boundaries?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = much less noisy, 5 = much more noisy)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

## Section 4 — Link to Survival Prediction

4.1 Do the AI–GT differences likely capture prognostically relevant tissue that improves survival prediction?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

4.2 How would you rate the effect of the AI’s deviation from GT (larger or smaller volume) on survival prediction?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = strongly harmful, 5 = strongly beneficial)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

## Section 5 — Trust & Preference (Survival Prediction Focus)

5.1 How likely is it that AI-generated masks are more accurate than GT in some cases (for example, subtle lesions missed by physicians, or over-delineation that reduces noise) in survival prediction tasks?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = very unlikely, 5 = very likely)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

5.2 Would you prefer to use the AI mask (rather than GT) as input for survival prediction?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not at all, 5 = strongly yes)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

## Section 6 — Summaries and Overall Assessment

6.1 Diagnosis-oriented evaluation — How successful is the AI mask in lesion delineation compared to GT?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = complete failure, 5 = highly successful)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6.2 Prediction-oriented evaluation — How successful is the AI mask in contributing to survival prediction compared to GT?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no added value, 5 = very strong added value)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6.3 Overall confidence in your assessment

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no confidence, 5 = complete confidence)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6.4 Do you think geometric evaluation metrics (e.g., Dice, IoU, Hausdorff distance) are sufficient to establish clinical trust in AI segmentation?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not sufficient at all, 5 = fully sufficient)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6.5 What additional types of evaluation do you think are necessary for trustworthy AI segmentation (e.g., radiomics stability, survival prediction, physician review, workflow impact, and others), and why do you think those are effective in building trustworthiness?

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

## Section 7 — Clinical Adoption & Workflow Impact

7.1 Usability — How easy is it to interpret and use the AI mask compared to GT in clinical workflow?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = much harder, 5 = much easier)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.2 Clinical actionability — Would the AI–GT differences change clinical decisions (for example, radiotherapy target, surgery extent, follow-up planning)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no impact, 5 = strong impact)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.3 Robustness and variability — How consistent and reliable are AI masks across different cases and observers?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = very inconsistent, 5 = very consistent)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.4 Time-saving potential — How much time could this AI save compared to manual delineation?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no time saved, 5 = saves most effort)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.5 Trust in clinical use — How confident would you be using AI masks for high-stakes tasks (e.g., radiotherapy planning) without physician refinement?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no trust, 5 = complete trust)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.6 Confidence maps/interpretability (if available) — Do uncertainty or confidence maps improve your trust in AI outputs?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no improvement, 5 = very strong improvement)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.7 Role of AI in clinical practice — In your opinion, AI-based segmentation should:

☐ Replace physicians entirely  
☐ Provide an initial mask for physicians to refine  
☐ Assist physicians as a support/check tool

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**End**

**Comparative Clinical Assessment of Deep Learning Tumor Masks vs. Ground Truth in Lung Cancer**

Model under review: ResUnet

**Segmentation Results**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Analysis Mode | Model | Dice Score | IoU Score | Hausdorff Distance |
| Test |  |  |  |  |
| **ResUNet** | 75 0.08 | 0.61 0.10 | 17.95 6.91 |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

**Prediction Result:**

Radiomics extracted from Ground Truth used for prediction tasks

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Validation | | | | | | Test | | | | | | | Names | | | |
| Acc | Precision | Recall | F1 | AUC | Specificity | Acc | Precision | Recall | F1 | AUC | Specificity | FSA/AEA | | CA | Model | SL/SSL |
| 0.88 | 0.78 | 0.88 | 0.83 | 0.54 | 1 | 0.87 | 0.76 | 0.87 | 0.81 | 0.71 | 1 | Mutual\_Info\_Gain\_Ratio | | ExtraTree | GT | SSL |
| 0.74 | 0.67 | 0.74 | 0.66 | 0.55 | 0.98 | 0.87 | 0.79 | 0.87 | 0.83 | 0.63 | 0.99 | SL |

Radiomics extracted from DL ResUNet used for prediction tasks

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Validation | | | | | | Test | | | | | | | Names | | | |
| Acc | Precision | Recall | F1 | AUC | Specificity | Acc | Precision | Recall | F1 | AUC | Specificity | FSA/AEA | | CA | Model | SL/SSL |
| 0.88 | 0.78 | 0.88 | 0.83 | 0.30 | 1 | 0.84 | 0.71 | 0.84 | 0.77 | 0.67 | 1 | Mutual Information | | MLP | ResUNet | SSL |
| 0.75 | 0.56 | 0.75 | 0.64 | 0.56 | 1 | 0.83 | 0.69 | 0.83 | 0.75 | 0.68 | 1 | SL |

**Start**

Section 1 — Overall Impression

1.1 Does the AI mask define a tumor volume that is more clinically meaningful than the Ground Truth (GT) for diagnosis and visual interpretation?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not at all, 5 = very strongly yes)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 2 — Spatial Differences

2.1 Are the AI–GT differences concentrated in the peritumoral zone (≈5–15 mm around the tumor)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not at all, 5 = strongly concentrated)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

2.2 How clinically meaningful are the regions where the AI mask differs from GT (including over-segmentation or under-segmentation, such as peritumoral expansion, missed necrotic cores, spiculated margins)? Or, how clinically justified are the AI–GT differences (whether due to additional regions or excluded regions)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 3 — Clinical Interpretation of Differences

3.1 How well does the AI mask preserve internal tumor heterogeneity (for example: solid vs. ground-glass opacity, necrotic vs. viable regions)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

3.2 Boundary quality: compared to GT, how noisy or fragmented are AI boundaries?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = much less noisy, 5 = much more noisy)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 4 — Link to Survival Prediction

4.1 Do the AI–GT differences likely capture prognostically relevant tissue that improves survival prediction?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

4.2 How would you rate the effect of the AI’s deviation from GT (larger or smaller volume) on survival prediction?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = strongly harmful, 5 = strongly beneficial)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 5 — Trust & Preference (Survival Prediction Focus)

5.1 How likely is it that AI-generated masks are more accurate than GT in some cases (for example, subtle lesions missed by physicians, or over-delineation that reduces noise) in survival prediction tasks?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = very unlikely, 5 = very likely)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

5.2 Would you prefer to use the AI mask (rather than GT) as input for survival prediction?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not at all, 5 = strongly yes)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 6 — Summaries and Overall Assessment

6.1 Diagnosis-oriented evaluation — How successful is the AI mask in lesion delineation compared to GT?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = complete failure, 5 = highly successful)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6.2 Prediction-oriented evaluation — How successful is the AI mask in contributing to survival prediction compared to GT?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no added value, 5 = very strong added value)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6.3 Overall confidence in your assessment

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no confidence, 5 = complete confidence)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6.4 Do you think geometric evaluation metrics (e.g., Dice, IoU, Hausdorff distance) are sufficient to establish clinical trust in AI segmentation?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not sufficient at all, 5 = fully sufficient)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6.5 What additional types of evaluation do you think are necessary for trustworthy AI segmentation (e.g., radiomics stability, survival prediction, physician review, workflow impact, and others), and why do you think those are effective in building trustworthiness?

Reason: In addition to evaluating the potential risks for each individual, I believe that if a consensus is reached between developers and physicians regarding the workflow for using images, the results can be more effectively integrated into clinical practice.

Section 7 — Clinical Adoption & Workflow Impact

7.1 Usability — How easy is it to interpret and use the AI mask compared to GT in clinical workflow?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = much harder, 5 = much easier)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.2 Clinical actionability — Would the AI–GT differences change clinical decisions (for example, radiotherapy target, surgery extent, follow-up planning)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no impact, 5 = strong impact)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.3 Robustness and variability — How consistent and reliable are AI masks across different cases and observers?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = very inconsistent, 5 = very consistent)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.4 Time-saving potential — How much time could this AI save compared to manual delineation?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no time saved, 5 = saves most effort)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.5 Trust in clinical use — How confident would you be using AI masks for high-stakes tasks (e.g., radiotherapy planning) without physician refinement?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no trust, 5 = complete trust)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.6 Confidence maps/interpretability (if available) — Do uncertainty or confidence maps improve your trust in AI outputs?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no improvement, 5 = very strong improvement)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.7 Role of AI in clinical practice — In your opinion, AI-based segmentation should:

☐ Replace physicians entirely  
☐ Provide an initial mask for physicians to refine  
☐ Assist physicians as a support/check tool

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**End**

**Comparative Clinical Assessment of Deep Learning Tumor Masks vs. Ground Truth in Lung Cancer**

Model under review: VNet

**Segmentation Results**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Analysis Mode | Model | Dice Score | IoU Score | Hausdorff Distance |
| Test | **VNet** | 0.81 0.08 | 0.69 0.11 | 11.12 3.73 |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

**Prediction Result:**

Radiomics extracted from Ground Truth used for prediction tasks

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Validation | | | | | | Test | | | | | | Names | | | | |
| Acc | Precision | Recall | F1 | AUC | Specificity | Acc | Precision | Recall | F1 | AUC | Specificity | | FSA/AEA | CA | Model | SL/SSL |
| 0.88 | 0.78 | 0.88 | 0.83 | 0.54 | 1 | 0.87 | 0.76 | 0.87 | 0.81 | 0.71 | 1 | | Mutual\_Info\_Gain\_Ratio | ExtraTree | GT | SSL |
| 0.74 | 0.67 | 0.74 | 0.66 | 0.55 | 0.98 | 0.87 | 0.79 | 0.87 | 0.83 | 0.63 | 0.99 | | SL |

Radiomics extracted from Vnet used for prediction tasks

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Validation | | | | | | Test | | | | | | Names | | | |
| Acc | Precision | Recall | F1 | AUC | Specificity | Acc | Precision | Recall | F1 | AUC | Specificity | FSA/AEA | CA | Model | SL/SSL |
| 0.88 | 0.78 | 0.88 | 0.83 | 0.65 | 1 | 0.88 | 0.77 | 0.88 | 0.82 | 0.76 | 1 | LASSO | LGBM | VNet | SSL |
| 0.74 | 0.55 | 0.74 | 0.63 | 0.58 | 0.99 | 0.87 | 0.77 | 0.87 | 0.81 | 0.40 | 0.99 | SL |

**Start**

Section 1 — Overall Impression

1.1 Does the AI mask define a tumor volume that is more clinically meaningful than the Ground Truth (GT) for diagnosis and visual interpretation?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not at all, 5 = very strongly yes)

Reason: In several cases, there is no clinically meaningful definition of tumor volume; however, in some instances, VNet performs well in defining clinically relevant tumor volumes.

Section 2 — Spatial Differences

2.1 Are the AI–GT differences concentrated in the peritumoral zone (≈5–15 mm around the tumor)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not at all, 5 = strongly concentrated)

Reason: In most cases, the differences between AI and GT cover the peritumoral zone.

2.2 How clinically meaningful are the regions where the AI mask differs from GT (including over-segmentation or under-segmentation, such as peritumoral expansion, missed necrotic cores, spiculated margins)? Or, how clinically justified are the AI–GT differences (whether due to additional regions or excluded regions)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

Reason: In my opinion, VNet provides better clinical coverage.

Section 3 — Clinical Interpretation of Differences

3.1 How well does the AI mask preserve internal tumor heterogeneity (for example: solid vs. ground-glass opacity, necrotic vs. viable regions)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

Reason: In most cases, internal tumor heterogeneity can be observed.

3.2 Boundary quality: compared to GT, how noisy or fragmented are AI boundaries?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = much less noisy, 5 = much more noisy)

Reason: VNet produces less noise.

Section 4 — Link to Survival Prediction

4.1 Do the AI–GT differences likely capture prognostically relevant tissue that improves survival prediction?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

Reason: In both cases—radiomics extracted from Ground Truth (GT) and radiomics extracted from VNet—the models show strong predictive performance for survival outcomes, with similar accuracy (Acc) and precision metrics, especially in the case of VNet.

4.2 How would you rate the effect of the AI’s deviation from GT (larger or smaller volume) on survival prediction?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = strongly harmful, 5 = strongly beneficial)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 5 — Trust & Preference (Survival Prediction Focus)

5.1 How likely is it that AI-generated masks are more accurate than GT in some cases (for example, subtle lesions missed by physicians, or over-delineation that reduces noise) in survival prediction tasks?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = very unlikely, 5 = very likely)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

5.2 Would you prefer to use the AI mask (rather than GT) as input for survival prediction?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not at all, 5 = strongly yes)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 6 — Summaries and Overall Assessment

6.1 Diagnosis-oriented evaluation — How successful is the AI mask in lesion delineation compared to GT?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = complete failure, 5 = highly successful)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6.2 Prediction-oriented evaluation — How successful is the AI mask in contributing to survival prediction compared to GT?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no added value, 5 = very strong added value)

Reason: Using AI, VNet helps cover more areas that are missed by GT masks.

6.3 Overall confidence in your assessment

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no confidence, 5 = complete confidence)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6.4 Do you think geometric evaluation metrics (e.g., Dice, IoU, Hausdorff distance) are sufficient to establish clinical trust in AI segmentation?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not sufficient at all, 5 = fully sufficient)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6.5 What additional types of evaluation do you think are necessary for trustworthy AI segmentation (e.g., radiomics stability, survival prediction, physician review, workflow impact, and others), and why do you think those are effective in building trustworthiness?

Reason: Evaluations should be conducted both at the individual case level and the group level to assess accuracy across various scales, thereby enhancing the trustworthiness of AI segmentation. An individual evaluation score or scale is needed to discuss risk and confidence.

Section 7 — Clinical Adoption & Workflow Impact

7.1 Usability — How easy is it to interpret and use the AI mask compared to GT in clinical workflow?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = much harder, 5 = much easier)

Reason: It is still unclear whether a model can be developed that addresses all aspects of lung cancer for every individual.

7.2 Clinical actionability — Would the AI–GT differences change clinical decisions (for example, radiotherapy target, surgery extent, follow-up planning)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no impact, 5 = strong impact)

Reason: Yes, in my opinion, in some cases, these differences could impact clinical decisions.

7.3 Robustness and variability — How consistent and reliable are AI masks across different cases and observers?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = very inconsistent, 5 = very consistent)

Reason:

7.4 Time-saving potential — How much time could this AI save compared to manual delineation?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no time saved, 5 = saves most effort)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.5 Trust in clinical use — How confident would you be using AI masks for high-stakes tasks (e.g., radiotherapy planning) without physician refinement?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no trust, 5 = complete trust)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.6 Confidence maps/interpretability (if available) — Do uncertainty or confidence maps improve your trust in AI outputs?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no improvement, 5 = very strong improvement)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.7 Role of AI in clinical practice — In your opinion, AI-based segmentation should:

☐ Replace physicians entirely  
☐ Provide an initial mask for physicians to refine  
☐ Assist physicians as a support/check tool

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**End**

**D4**

**Comparative Clinical Assessment of Deep Learning Tumor Masks vs. Ground Truth in Lung Cancer**

Model under review: 3DAttention

**Segmentation Results**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Analysis Mode | Model | Dice Score | IoU Score | Hausdorff Distance |
| Test |  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
| **3DAttention** | 0.83 0.04 | 0.72 0.04 | 12.12 3.73 |

**Prediction Result:**

Radiomics extracted from Ground Truth used for prediction tasks

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Validation | | | | | | Test | | | | | | | Names | | | |
| Acc | Precision | Recall | F1 | AUC | Specificity | Acc | Precision | Recall | F1 | AUC | Specificity | FSA/AEA | | CA | Model | SL/SSL |
| 0.88 | 0.78 | 0.88 | 0.83 | 0.54 | 1 | 0.87 | 0.76 | 0.87 | 0.81 | 0.71 | 1 | Mutual\_Info\_Gain\_Ratio | | ExtraTree | GT | SSL |
| 0.74 | 0.67 | 0.74 | 0.66 | 0.55 | 0.98 | 0.87 | 0.79 | 0.87 | 0.83 | 0.63 | 0.99 | SL |

Radiomics extracted from DL 3D attention used for prediction tasks

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Validation | | | | | | Test | | | | | | | Names | | | |
| Acc | Precision | Recall | F1 | AUC | Specificity | Acc | Precision | Recall | F1 | AUC | Specificity | FSA/AEA | | CA | Model | SL/SSL |
| 0.79 | 0.80 | 0.79 | 0.80 | 0.69 | 0.87 | 0.81 | 0.83 | 0.81 | 0.82 | 0.76 | 0.88 | Guassian Random Projection | | Voting Classifier | 3DAtt | SSL |
| 0.61 | 0.64 | 0.61 | 0.62 | 0.53 | 0.69 | 0.72 | 0.79 | 0.72 | 0.75 | 0.55 | 0.79 | SL |

**Start**

Section 1 — Overall Impression

1.1 Does the AI mask define a tumor volume that is more clinically meaningful than the Ground Truth (GT) for diagnosis and visual interpretation?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 \* 5 ☐ (0 = not at all, 5 = very strongly yes)

Reason: It is likely that, to speed up ground truth (GT) segmentation, a high-contrast setting was used, so many of the GTs only include the highest-intensity regions. In contrast, the 3DAttention segmentation appears as a the biggest version of the GT, covering a larger peritumoral area. The major difference that seems not to be helping is a lot of noise around the main segmentation. Although there are some areas of oversegmentation and undersegmentation, it is less than Vnet.

Section 2 — Spatial Differences

2.1 Are the AI–GT differences concentrated in the peritumoral zone (≈5–15 mm around the tumor)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 \* (0 = not at all, 5 = strongly concentrated)

Reason: For most of the parts and images, this is true.

2.2 How clinically meaningful are the regions where the AI mask differs from GT (including over-segmentation or under-segmentation, such as peritumoral expansion, missed necrotic cores, spiculated margins)? Or, how clinically justified are the AI–GT differences (whether due to additional regions or excluded regions)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 \* 5 ☐

Reason: It depends on the context. For example, if we are predicting a patient’s height, the data would not be meaningful. However, if the outcome of interest is prognosis, the data becomes clinically meaningful and justifiable. For instance, peritumoral areas have been shown to be important for prognosis and for predicting future tumor invasion.

Section 3 — Clinical Interpretation of Differences

3.1 How well does the AI mask preserve internal tumor heterogeneity (for example: solid vs. ground-glass opacity, necrotic vs. viable regions)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 \* 5 ☐

Reason: Both masks preserve heterogeneity, but in the 3DAttention mask, relatively larger ROI captures more heterogeneity. Sometimes it highlights regions that appear irrelevant, likely because it includes non-tumoral areas, but at other times, it identifies less dense regions within the tumor that are still important for our analysis.

3.2 Boundary quality: compared to GT, how noisy or fragmented are AI boundaries?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 \* (0 = much less noisy, 5 = much more noisy)

Reason: AI contours are definitely more fragmented and noisier.

Section 4 — Link to Survival Prediction

4.1 Do the AI–GT differences likely capture prognostically relevant tissue that improves survival prediction?

0 ☐ 1 ☐ 2 ☐ 3 \*4 ☐ 5 ☐

Reason: Yes, in terms of both heterogeneity and peritumoral areas, this can help in survival prediction. However, the larger ROIs might introduce a bias toward predicting worse survival outcomes. Also, the amount of noise may distract the final model from predicting the survival outcome properly.

4.2 How would you rate the effect of the AI’s deviation from GT (larger or smaller volume) on survival prediction?

0 ☐ 1 ☐ 2 \* 3 ☐ 4 ☐ 5 ☐ (0 = strongly harmful, 5 = strongly beneficial)

Reason: I’m on the fence about this. While it may be beneficial in some cases, in others it fails to accurately capture the ROI or produces an unreasonably large ROI. Also, there is a lot of noise in the segmentation, which mostly seems to be non-informative

Section 5 — Trust & Preference (Survival Prediction Focus)

5.1 How likely is it that AI-generated masks are more accurate than GT in some cases (for example, subtle lesions missed by physicians, or over-delineation that reduces noise) in survival prediction tasks?

0 ☐ 1 ☐ 2 \* 3 ☐ 4 ☐ 5 ☐ (0 = very unlikely, 5 = very likely)

Reason: Although it performs better than the GT in capturing peritumoral areas and heterogeneity, if it fails to identify some important tumor regions, it could negatively impact patient care. The size of the active tumor remains one of the most important features, particularly for survival prediction. Also, the amount of noise reduces trust.

5.2 Would you prefer to use the AI mask (rather than GT) as input for survival prediction?

0 ☐ 1 ☐ 2 \* 3 ☐ 4 ☐ 5 ☐ (0 = not at all, 5 = strongly yes)

Reason: I would if there is some explanation of how some very important areas are not included in ROI, and even providing different contrast and settings of the image so that I might be able to see some hidden features that the AI segmented the image based on.

Section 6 — Summaries and Overall Assessment

6.1 Diagnosis-oriented evaluation — How successful is the AI mask in lesion delineation compared to GT?

0 ☐ 1 ☐ 2 ☐ 3 \* 4 ☐ 5 ☐ (0 = complete failure, 5 = highly successful)

Reason: The reasons and facts are discussed above

6.2 Prediction-oriented evaluation — How successful is the AI mask in contributing to survival prediction compared to GT?

0 ☐ 1 ☐ 2 \* 3 ☐ 4 ☐ 5 ☐ (0 = no added value, 5 = very strong added value)

Reason: the reasons and facts are discussed above

6.3 Overall confidence in your assessment

0 ☐ 1 ☐ 2 ☐ 3 \* 4 ☐ 5 ☐ (0 = no confidence, 5 = complete confidence)

Reason: Given that this task ideally requires input from radiologists and oncologists, I am not entirely confident in my assessment.

6.4 Do you think geometric evaluation metrics (e.g., Dice, IoU, Hausdorff distance) are sufficient to establish clinical trust in AI segmentation?

0 ☐ 1 \* 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not sufficient at all, 5 = fully sufficient)

Reason: Actually, the results show the opposite. The key difference is that these AI masks perform better overall. While they can be helpful in certain aspects, such as assessing tumor size, there are still many shortcomings in using these methods.

6.5 What additional types of evaluation do you think are necessary for trustworthy AI segmentation (e.g., radiomics stability, survival prediction, physician review, workflow impact, and others), and why do you think those are effective in building trustworthiness?

Reason: I think radiomics stability still suffers from the same pitfalls as the Dice score. Therefore, focusing on the outcome of interest may often be a better option, although using workflow impact is preferable to relying solely on outcome prediction.

Section 7 — Clinical Adoption & Workflow Impact

7.1 Usability — How easy is it to interpret and use the AI mask compared to GT in clinical workflow?

0 ☐ 1 ☐ 2 \* 3 ☐ 4 ☐ 5 ☐ (0 = much harder, 5 = much easier)

Reason: Although it contains more pretumoral area and it is more integrated, making it easier to follow and include features, it still has major failures in certain areas, such as including non-tumoral regions or excluding parts of the tumor. Reducing or justifying excessive noise should also be considered.

7.2 Clinical actionability — Would the AI–GT differences change clinical decisions (for example, radiotherapy target, surgery extent, follow-up planning)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 \* (0 = no impact, 5 = strong impact)

Reason: For sure, specialists rely heavily on the ROI, especially for non-invasive procedures like radiotherapy. Even small changes can affect both the size and location of the tumor. Additionally, under-segmentation may cause the model to miss tumors altogether.

7.3 Robustness and variability — How consistent and reliable are AI masks across different cases and observers?

0 ☐ 1 \* 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = very inconsistent, 5 = very consistent)

Reason: Some features remained the same, but for most, the criteria for distinguishing tumoral from non-tumoral areas differed.

7.4 Time-saving potential — How much time could this AI save compared to manual delineation?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 \* 5 ☐ (0 = no time saved, 5 = saves most effort)

Reason: When implemented correctly and supervised to address potential failures, it can be very helpful.

7.5 Trust in clinical use — How confident would you be using AI masks for high-stakes tasks (e.g., radiotherapy planning) without physician refinement?

0 ☐ 1 \* 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no trust, 5 = complete trust)

Reason: the reasons and facts are discussed above

7.6 Confidence maps/interpretability (if available) — Do uncertainty or confidence maps improve your trust in AI outputs?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 \* (0 = no improvement, 5 = very strong improvement)

Reason: as I mentioned ealier this is very helpful and can improve trust easily.

7.7 Role of AI in clinical practice — In your opinion, AI-based segmentation should:

☐ Replace physicians entirely  
☐ Provide an initial mask for physicians to refine  
\* Assist physicians as a support/check tool

Reason: For now, and likely for the foreseeable future, human knowledge is not extensive enough for algorithms to learn every possible source of noise or anomaly. Even physicians regularly encounter difficult and novel cases. Until human knowledge—or sufficiently diverse datasets—reaches that level, these algorithms should remain as assistants rather than autonomous decision-makers.

**End**

**Comparative Clinical Assessment of Deep Learning Tumor Masks vs. Ground Truth in Lung Cancer**

Model under review: 3DMedSam

**Segmentation Results**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Analysis Mode | Model | Dice Score | IoU Score | Hausdorff Distance |
| Test |  |  |  |  |
|  |  |  |  |
|  |  |  |  |
| **3DMedSam** | 0.68 0.19 | 0.52 0.18 | 4.26 1.55 |

**Prediction Result:**

Radiomics extracted from Ground Truth used for prediction tasks

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Validation | | | | | | Test | | | | | | | Names | | | |
| Acc | Precision | Recall | F1 | AUC | Specificity | Acc | Precision | Recall | F1 | AUC | Specificity | FSA/AEA | | CA | Model | SL/SSL |
| 0.88 | 0.78 | 0.88 | 0.83 | 0.54 | 1 | 0.87 | 0.76 | 0.87 | 0.81 | 0.71 | 1 | Mutual\_Info\_Gain\_Ratio | | ExtraTree | GT | SSL |
| 0.74 | 0.67 | 0.74 | 0.66 | 0.55 | 0.98 | 0.87 | 0.79 | 0.87 | 0.83 | 0.63 | 0.99 | SL |

Radiomics extracted from DL 3DmedSam used for prediction tasks

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Validation | | | | | | Test | | | | | | Names | | | |
| Acc | Precision | Recall | F1 | AUC | Specificity | Acc | Precision | Recall | F1 | AUC | Specificity | FSA/AEA | CA | Model | SL/SSL |
| 0.88 | 0.77 | 0.88 | 0.82 | 0.67 | 1 | 0.85 | 0.71 | 0.85 | 0.77 | 0.68 | 1 | ReliefF | GradientBoosting | MedSam | SSL |
| 0.74 | 0.55 | 0.74 | 0.63 | 0.51 | 0.99 | 0.86 | 0.76 | 0.86 | 0.81 | 0.71 | 0.99 | SL |

**Start**

Section 1 — Overall Impression

1.1 Does the AI mask define a tumor volume that is more clinically meaningful than the Ground Truth (GT) for diagnosis and visual interpretation?

0 ☐ 1 \* 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not at all, 5 = very strongly yes)

Reason: It is likely that, to speed up ground truth (GT) segmentation, a high-contrast setting was used, so many of the GTs only include the highest-intensity regions. Among the models, 3DMedSAM shows the most similarity to the GT segmentation, particularly in producing less fragmented contours. At times, the algorithm performs very well in detecting features and tumor GTV regions that no other method captures, but more often it under-segments important areas. Compared to other approaches, it also includes fewer peritumoral regions.

Section 2 — Spatial Differences

2.1 Are the AI–GT differences concentrated in the peritumoral zone (≈5–15 mm around the tumor)?

0 \* 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not at all, 5 = strongly concentrated)

Reason: The ROIs generally do not include any part of the peritumoral region.

2.2 How clinically meaningful are the regions where the AI mask differs from GT (including over-segmentation or under-segmentation, such as peritumoral expansion, missed necrotic cores, spiculated margins)? Or, how clinically justified are the AI–GT differences (whether due to additional regions or excluded regions)?

0 ☐ 1 \* 2 ☐ 3 ☐ 4 ☐ 5 ☐

Reason: It depends on the context. For example, if we were predicting a patient’s height, the data would not be meaningful. However, if the outcome of interest is prognosis, it is unlikely that this segmentation captures more than the simple GT. Moreover, due to the high number of false-positive regions, it often introduces misleading results.

Section 3 — Clinical Interpretation of Differences

3.1 How well does the AI mask preserve internal tumor heterogeneity (for example: solid vs. ground-glass opacity, necrotic vs. viable regions)?

0 ☐ 1 \*2 ☐ 3 ☐ 4 ☐ 5 ☐

Reason: 3DMedSAM mainly includes areas of the primary GTV, seemingly focusing on regions with the highest contrast and other regions within the ROI without a clear rationale. As a result, the level of heterogeneity in this segmentation is very low.

3.2 Boundary quality: compared to GT, how noisy or fragmented are AI boundaries?

0 ☐ 1 ☐ 2 \*3 ☐ 4 ☐ 5 (0 = much less noisy, 5 = much more noisy)

Reason: It has almost the same noise in case of bounderies.

Section 4 — Link to Survival Prediction

4.1 Do the AI–GT differences likely capture prognostically relevant tissue that improves survival prediction?

0 ☐ 1 ☐ 2 ☐ 3 \* 4 ☐ 5 ☐

Reason: The only clear advantage of this segmentation is its inclusion of more GTV compared to the GT segmentation, which can be very helpful for prognosis.

4.2 How would you rate the effect of the AI’s deviation from GT (larger or smaller volume) on survival prediction?

0 ☐ 1 ☐ 2 ☐ 3 \* 4 ☐ 5 ☐ (0 = strongly harmful, 5 = strongly beneficial)

Reason: I’m on the fence about this. While it may be beneficial in some cases, in others it fails to accurately capture the ROI or produces an unreasonably large ROI.

Section 5 — Trust & Preference (Survival Prediction Focus)

5.1 How likely is it that AI-generated masks are more accurate than GT in some cases (for example, subtle lesions missed by physicians, or over-delineation that reduces noise) in survival prediction tasks?

0 ☐ 1 \* 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = very unlikely, 5 = very likely)

Reason: in some cases AI can have better acuuracy for tumor but in most cases it is not like this.

5.2 Would you prefer to use the AI mask (rather than GT) as input for survival prediction?

0 ☐ 1 \* 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not at all, 5 = strongly yes)

Reason: I would if there is some explanation of how some very important areas are not included in ROI.

Section 6 — Summaries and Overall Assessment

6.1 Diagnosis-oriented evaluation — How successful is the AI mask in lesion delineation compared to GT?

0 ☐ 1 \* 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = complete failure, 5 = highly successful)

Reason: the reasons and facts are discussed above

6.2 Prediction-oriented evaluation — How successful is the AI mask in contributing to survival prediction compared to GT?

0 ☐ 1 \* 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no added value, 5 = very strong added value)

Reason: the reasons and facts are discussed above

6.3 Overall confidence in your assessment

0 ☐ 1 ☐ 2 ☐ 3 \* 4 ☐ 5 ☐ (0 = no confidence, 5 = complete confidence)

Reason: Given that this task ideally requires input from radiologists and oncologists, I am not entirely confident in my assessment.

6.4 Do you think geometric evaluation metrics (e.g., Dice, IoU, Hausdorff distance) are sufficient to establish clinical trust in AI segmentation?

0 ☐ 1 \* 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not sufficient at all, 5 = fully sufficient)

Reason: Actually, the results show the opposite. The key difference is that these AI masks perform better overall. While they can be helpful in certain aspects, such as assessing tumor size, there are still many shortcomings in using these methods.

6.5 What additional types of evaluation do you think are necessary for trustworthy AI segmentation (e.g., radiomics stability, survival prediction, physician review, workflow impact, and others), and why do you think those are effective in building trustworthiness?

Reason: I think radiomics stability still suffers from the same pitfalls as the Dice score. Therefore, focusing on the outcome of interest may often be a better option, although using workflow impact is preferable to relying solely on outcome prediction.

Section 7 — Clinical Adoption & Workflow Impact

7.1 Usability — How easy is it to interpret and use the AI mask compared to GT in clinical workflow?

0 ☐ 1 \* 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = much harder, 5 = much easier)

Reason: Although it is more integrated, making it easier to follow and capture features, it still has major shortcomings, such as including non-tumoral regions or excluding important parts of the tumor.

7.2 Clinical actionability — Would the AI–GT differences change clinical decisions (for example, radiotherapy target, surgery extent, follow-up planning)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 \* (0 = no impact, 5 = strong impact)

Reason: For sure, specialists rely heavily on the ROI, especially for non-invasive procedures like radiotherapy. Even small changes can affect both the size and location of the tumor. Additionally, under-segmentation may cause the model to miss tumors altogether.

7.3 Robustness and variability — How consistent and reliable are AI masks across different cases and observers?

0 \* 1 ☐2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = very inconsistent, 5 = very consistent)

Reason: Some features remained the same, but for most, the criteria for distinguishing tumoral from non-tumoral areas differed.

7.4 Time-saving potential — How much time could this AI save compared to manual delineation?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 \* 5 ☐ (0 = no time saved, 5 = saves most effort)

Reason: When implemented correctly and supervised to address potential failures, it can be very helpful.

7.5 Trust in clinical use — How confident would you be using AI masks for high-stakes tasks (e.g., radiotherapy planning) without physician refinement?

0 ☐ 1 \* 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no trust, 5 = complete trust)

Reason: the reasons and facts are discussed above

7.6 Confidence maps/interpretability (if available) — Do uncertainty or confidence maps improve your trust in AI outputs?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 \* (0 = no improvement, 5 = very strong improvement)

Reason: as I mentioned ealier this Is very helpful and can improve trust easily.

7.7 Role of AI in clinical practice — In your opinion, AI-based segmentation should:

☐ Replace physicians entirely  
☐ Provide an initial mask for physicians to refine  
\* Assist physicians as a support/check tool

Reason: For now, and likely for the foreseeable future, human knowledge is not extensive enough for algorithms to learn every possible source of noise or anomaly. Even physicians regularly encounter difficult and novel cases. Until human knowledge—or sufficiently diverse datasets—reaches that level, these algorithms should remain as assistants rather than autonomous decision-makers.

**End**

**Comparative Clinical Assessment of Deep Learning Tumor Masks vs. Ground Truth in Lung Cancer**

Model under review: Reconnet

**Segmentation Results**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Analysis Mode | Model | Dice Score | IoU Score | Hausdorff Distance |
| Test |  |  |  |  |
|  |  |  |  |
| **Reconnet** | 0.80 0.09 | 0.68 0.13 | 11.30 3.96 |

**Prediction Result:**

Radiomics extracted from Ground Truth used for prediction tasks

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Validation | | | | | | Test | | | | | | | Names | | | |
| Acc | Precision | Recall | F1 | AUC | Specificity | Acc | Precision | Recall | F1 | AUC | Specificity | FSA/AEA | | CA | Model | SL/SSL |
| 0.88 | 0.78 | 0.88 | 0.83 | 0.54 | 1 | 0.87 | 0.76 | 0.87 | 0.81 | 0.71 | 1 | Mutual\_Info\_Gain\_Ratio | | ExtraTree | GT | SSL |
| 0.74 | 0.67 | 0.74 | 0.66 | 0.55 | 0.98 | 0.87 | 0.79 | 0.87 | 0.83 | 0.63 | 0.99 | SL |

Radiomics extracted from DL Reconnet used for prediction tasks

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Validation | | | | | | Test | | | | | | Names | | | |
| Acc | Precision | Recall | F1 | AUC | Specificity | Acc | Precision | Recall | F1 | AUC | Specificity | FSA/AEA | CA | Model | SL/SSL |
| 0.88 | 0.78 | 0.88 | 0.83 | 0.65 | 1 | 0.86 | 0.75 | 0.86 | 0.80 | 0.71 | 1 | UMAP | AdaBoost | Reconnet | SSL |
| 0.75 | 0.56 | 0.75 | 0.64 | 0.50 | 1 | 0.88 | 0.77 | 0.88 | 0.82 | 0.5 | 1 | SL |

**Start**

Section 1 — Overall Impression

1.1 Does the AI mask define a tumor volume that is more clinically meaningful than the Ground Truth (GT) for diagnosis and visual interpretation?

0 ☐ 1 ☐ 2 ☐ 3 \* 4 ☐ 5 ☐ (0 = not at all, 5 = very strongly yes)

Reason: It is likely that, to speed up ground truth (GT) segmentation, a high-contrast setting was used, so many of the GTs only include the highest-intensity regions. In contrast, the ResUnet segmentation appears as a smoother and bigger version of the GT, covering a larger peritumoral area. Occasionally, it fails to detect certain features and regions that are important, but sometimes it identifies areas that the GT did not include, which can be significant for diagnosis and interpretation.

Section 2 — Spatial Differences

2.1 Are the AI–GT differences concentrated in the peritumoral zone (≈5–15 mm around the tumor)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 \* 5 ☐ (0 = not at all, 5 = strongly concentrated)

Reason: For most of the parts and images this is true.

2.2 How clinically meaningful are the regions where the AI mask differs from GT (including over-segmentation or under-segmentation, such as peritumoral expansion, missed necrotic cores, spiculated margins)? Or, how clinically justified are the AI–GT differences (whether due to additional regions or excluded regions)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 \* 5 ☐

Reason: It depends on the context. For example, if we are predicting a patient’s height, the data would not be meaningful. However, if the outcome of interest is prognosis, the data becomes clinically meaningful and justifiable. For instance, peritumoral areas have been shown to be important for prognosis and for predicting future tumor invasion.

Section 3 — Clinical Interpretation of Differences

3.1 How well does the AI mask preserve internal tumor heterogeneity (for example: solid vs. ground-glass opacity, necrotic vs. viable regions)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 \* 5 ☐

Reason: Both masks preserve heterogeneity, but in the ResUnet mask, the smoother and relatively larger ROI captures more heterogeneity. Sometimes it highlights regions that appear irrelevant, likely because it includes non-tumoral areas, but at other times, it identifies less dense regions within the tumor that are still important for our analysis.

3.2 Boundary quality: compared to GT, how noisy or fragmented are AI boundaries?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 \* (0 = much less noisy, 5 = much more noisy)

Reason: AI contours are definitely more fragmented and as the result smoother.

Section 4 — Link to Survival Prediction

4.1 Do the AI–GT differences likely capture prognostically relevant tissue that improves survival prediction?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 \* 5 ☐

Reason: Yes, in terms of both heterogeneity and peritumoral areas, this can help in survival prediction. However, the larger ROIs might introduce a bias toward predicting worse survival outcomes.

4.2 How would you rate the effect of the AI’s deviation from GT (larger or smaller volume) on survival prediction?

0 ☐ 1 ☐ 2 ☐ 3 \* 4 ☐ 5 ☐ (0 = strongly harmful, 5 = strongly beneficial)

Reason: I’m on the fence about this. While it may be beneficial in some cases, in others it fails to accurately capture the ROI or produces an unreasonably large ROI.

Section 5 — Trust & Preference (Survival Prediction Focus)

5.1 How likely is it that AI-generated masks are more accurate than GT in some cases (for example, subtle lesions missed by physicians, or over-delineation that reduces noise) in survival prediction tasks?

0 ☐ 1 ☐ 2 ☐ 3 \* 4 ☐ 5 ☐ (0 = very unlikely, 5 = very likely)

Reason: Although it performs better than the GT in capturing peritumoral areas and heterogeneity, if it fails to identify some important tumor regions, it could negatively impact patient care. The size of the active tumor remains one of the most important features, particularly for survival prediction.

5.2 Would you prefer to use the AI mask (rather than GT) as input for survival prediction?

0 ☐ 1 ☐ 2 \* 3 ☐ 4 ☐ 5 ☐ (0 = not at all, 5 = strongly yes)

Reason: I would if there is some explanation of how some very important areas are not included in ROI.

Section 6 — Summaries and Overall Assessment

6.1 Diagnosis-oriented evaluation — How successful is the AI mask in lesion delineation compared to GT?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 \* 5 ☐ (0 = complete failure, 5 = highly successful)

Reason: the reasons and facts are discussed above

6.2 Prediction-oriented evaluation — How successful is the AI mask in contributing to survival prediction compared to GT?

0 ☐ 1 ☐ 2 ☐ 3 \* 4 ☐ 5 ☐ (0 = no added value, 5 = very strong added value)

Reason: the reasons and facts are discussed above

6.3 Overall confidence in your assessment

0 ☐ 1 ☐ 2 ☐ 3 \* 4 ☐ 5 ☐ (0 = no confidence, 5 = complete confidence)

Reason: Given that this task ideally requires input from radiologists and oncologists, I am not entirely confident in my assessment.

6.4 Do you think geometric evaluation metrics (e.g., Dice, IoU, Hausdorff distance) are sufficient to establish clinical trust in AI segmentation?

0 ☐ 1 \* 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not sufficient at all, 5 = fully sufficient)

Reason: Actually, the results show the opposite. The key difference is that these AI masks perform better overall. While they can be helpful in certain aspects, such as assessing tumor size, there are still many shortcomings in using these methods.

6.5 What additional types of evaluation do you think are necessary for trustworthy AI segmentation (e.g., radiomics stability, survival prediction, physician review, workflow impact, and others), and why do you think those are effective in building trustworthiness?

Reason: I think radiomics stability still suffers from the same pitfalls as the Dice score. Therefore, focusing on the outcome of interest may often be a better option, although using workflow impact is preferable to relying solely on outcome prediction.

Section 7 — Clinical Adoption & Workflow Impact

7.1 Usability — How easy is it to interpret and use the AI mask compared to GT in clinical workflow?

0 ☐ 1 ☐ 2 ☐ 3 \* 4 ☐ 5 ☐ (0 = much harder, 5 = much easier)

Reason: Although it is smoother and more integrated, making it easier to follow and include features, it still has major failures in certain areas, such as including non-tumoral regions or excluding parts of the tumor.

7.2 Clinical actionability — Would the AI–GT differences change clinical decisions (for example, radiotherapy target, surgery extent, follow-up planning)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 \* (0 = no impact, 5 = strong impact)

Reason: For sure, specialists rely heavily on the ROI, especially for non-invasive procedures like radiotherapy. Even small changes can affect both the size and location of the tumor. Additionally, under-segmentation may cause the model to miss tumors altogether.

7.3 Robustness and variability — How consistent and reliable are AI masks across different cases and observers?

0 ☐ 1 \* 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = very inconsistent, 5 = very consistent)

Reason: Some features remained the same, but for most, the criteria for distinguishing tumoral from non-tumoral areas differed.

7.4 Time-saving potential — How much time could this AI save compared to manual delineation?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 \* 5 ☐ (0 = no time saved, 5 = saves most effort)

Reason: When implemented correctly and supervised to address potential failures, it can be very helpful.

7.5 Trust in clinical use — How confident would you be using AI masks for high-stakes tasks (e.g., radiotherapy planning) without physician refinement?

0 ☐ 1 \* 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no trust, 5 = complete trust)

Reason: the reasons and facts are discussed above

7.6 Confidence maps/interpretability (if available) — Do uncertainty or confidence maps improve your trust in AI outputs?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 \* (0 = no improvement, 5 = very strong improvement)

Reason: as I mentioned ealier this Is very helpful and can improve trust easily.

7.7 Role of AI in clinical practice — In your opinion, AI-based segmentation should:

☐ Replace physicians entirely  
☐ Provide an initial mask for physicians to refine  
\* Assist physicians as a support/check tool

Reason: For now, and likely for the foreseeable future, human knowledge is not extensive enough for algorithms to learn every possible source of noise or anomaly. Even physicians regularly encounter difficult and novel cases. Until human knowledge—or sufficiently diverse datasets—reaches that level, these algorithms should remain as assistants rather than autonomous decision-makers.

**End**

**Comparative Clinical Assessment of Deep Learning Tumor Masks vs. Ground Truth in Lung Cancer**

Model under review: ResUnet

**Segmentation Results**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Analysis Mode | Model | Dice Score | IoU Score | Hausdorff Distance |
| Test |  |  |  |  |
| **ResUNet** | 75 0.08 | 0.61 0.10 | 17.95 6.91 |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

**Prediction Result:**

Radiomics extracted from Ground Truth used for prediction tasks

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Validation | | | | | | Test | | | | | | | Names | | | |
| Acc | Precision | Recall | F1 | AUC | Specificity | Acc | Precision | Recall | F1 | AUC | Specificity | FSA/AEA | | CA | Model | SL/SSL |
| 0.88 | 0.78 | 0.88 | 0.83 | 0.54 | 1 | 0.87 | 0.76 | 0.87 | 0.81 | 0.71 | 1 | Mutual\_Info\_Gain\_Ratio | | ExtraTree | GT | SSL |
| 0.74 | 0.67 | 0.74 | 0.66 | 0.55 | 0.98 | 0.87 | 0.79 | 0.87 | 0.83 | 0.63 | 0.99 | SL |

Radiomics extracted from DL ResUNet used for prediction tasks

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Validation | | | | | | Test | | | | | | | Names | | | |
| Acc | Precision | Recall | F1 | AUC | Specificity | Acc | Precision | Recall | F1 | AUC | Specificity | FSA/AEA | | CA | Model | SL/SSL |
| 0.88 | 0.78 | 0.88 | 0.83 | 0.30 | 1 | 0.84 | 0.71 | 0.84 | 0.77 | 0.67 | 1 | Mutual Information | | MLP | ResUNet | SSL |
| 0.75 | 0.56 | 0.75 | 0.64 | 0.56 | 1 | 0.83 | 0.69 | 0.83 | 0.75 | 0.68 | 1 | SL |

**Start**

Section 1 — Overall Impression

1.1 Does the AI mask define a tumor volume that is more clinically meaningful than the Ground Truth (GT) for diagnosis and visual interpretation?

0 ☐ 1 ☐ 2 ☐ 3 \* 4 ☐ 5 ☐ (0 = not at all, 5 = very strongly yes)

Reason: It is likely that, to speed up ground truth (GT) segmentation, a high-contrast setting was used, so many of the GTs only include the highest-intensity regions. In contrast, the ResUnet segmentation appears as a smoother and bigger version of the GT, covering a larger peritumoral area. Occasionally, it fails to detect certain features and regions that are important, but sometimes it identifies areas that the GT did not include, which can be significant for diagnosis and interpretation.

Section 2 — Spatial Differences

2.1 Are the AI–GT differences concentrated in the peritumoral zone (≈5–15 mm around the tumor)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 \* 5 ☐ (0 = not at all, 5 = strongly concentrated)

Reason: For most of the parts and images this is true.

2.2 How clinically meaningful are the regions where the AI mask differs from GT (including over-segmentation or under-segmentation, such as peritumoral expansion, missed necrotic cores, spiculated margins)? Or, how clinically justified are the AI–GT differences (whether due to additional regions or excluded regions)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 \* 5 ☐

Reason: It depends on the context. For example, if we are predicting a patient’s height, the data would not be meaningful. However, if the outcome of interest is prognosis, the data becomes clinically meaningful and justifiable. For instance, peritumoral areas have been shown to be important for prognosis and for predicting future tumor invasion.

Section 3 — Clinical Interpretation of Differences

3.1 How well does the AI mask preserve internal tumor heterogeneity (for example: solid vs. ground-glass opacity, necrotic vs. viable regions)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 \* 5 ☐

Reason: Both masks preserve heterogeneity, but in the ResUnet mask, the smoother and relatively larger ROI captures more heterogeneity. Sometimes it highlights regions that appear irrelevant, likely because it includes non-tumoral areas, but at other times, it identifies less dense regions within the tumor that are still important for our analysis.

3.2 Boundary quality: compared to GT, how noisy or fragmented are AI boundaries?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 \* (0 = much less noisy, 5 = much more noisy)

Reason: AI contours are definitely more fragmented and as the result smoother.

Section 4 — Link to Survival Prediction

4.1 Do the AI–GT differences likely capture prognostically relevant tissue that improves survival prediction?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 \* 5 ☐

Reason: Yes, in terms of both heterogeneity and peritumoral areas, this can help in survival prediction. However, the larger ROIs might introduce a bias toward predicting worse survival outcomes.

4.2 How would you rate the effect of the AI’s deviation from GT (larger or smaller volume) on survival prediction?

0 ☐ 1 ☐ 2 ☐ 3 \* 4 ☐ 5 ☐ (0 = strongly harmful, 5 = strongly beneficial)

Reason: I’m on the fence about this. While it may be beneficial in some cases, in others it fails to accurately capture the ROI or produces an unreasonably large ROI.

Section 5 — Trust & Preference (Survival Prediction Focus)

5.1 How likely is it that AI-generated masks are more accurate than GT in some cases (for example, subtle lesions missed by physicians, or over-delineation that reduces noise) in survival prediction tasks?

0 ☐ 1 ☐ 2 ☐ 3 \* 4 ☐ 5 ☐ (0 = very unlikely, 5 = very likely)

Reason: Although it performs better than the GT in capturing peritumoral areas and heterogeneity, if it fails to identify some important tumor regions, it could negatively impact patient care. The size of the active tumor remains one of the most important features, particularly for survival prediction.

5.2 Would you prefer to use the AI mask (rather than GT) as input for survival prediction?

0 ☐ 1 ☐ 2 \* 3 ☐ 4 ☐ 5 ☐ (0 = not at all, 5 = strongly yes)

Reason: I would if there is some explanation of how some very important areas are not included in ROI.

Section 6 — Summaries and Overall Assessment

6.1 Diagnosis-oriented evaluation — How successful is the AI mask in lesion delineation compared to GT?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 \* 5 ☐ (0 = complete failure, 5 = highly successful)

Reason: the reasons and facts are discussed above

6.2 Prediction-oriented evaluation — How successful is the AI mask in contributing to survival prediction compared to GT?

0 ☐ 1 ☐ 2 ☐ 3 \* 4 ☐ 5 ☐ (0 = no added value, 5 = very strong added value)

Reason: the reasons and facts are discussed above

6.3 Overall confidence in your assessment

0 ☐ 1 ☐ 2 ☐ 3 \* 4 ☐ 5 ☐ (0 = no confidence, 5 = complete confidence)

Reason: Given that this task ideally requires input from radiologists and oncologists, I am not entirely confident in my assessment.

6.4 Do you think geometric evaluation metrics (e.g., Dice, IoU, Hausdorff distance) are sufficient to establish clinical trust in AI segmentation?

0 ☐ 1 \* 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not sufficient at all, 5 = fully sufficient)

Reason: Actually, the results show the opposite. The key difference is that these AI masks perform better overall. While they can be helpful in certain aspects, such as assessing tumor size, there are still many shortcomings in using these methods.

6.5 What additional types of evaluation do you think are necessary for trustworthy AI segmentation (e.g., radiomics stability, survival prediction, physician review, workflow impact, and others), and why do you think those are effective in building trustworthiness?

Reason: I think radiomics stability still suffers from the same pitfalls as the Dice score. Therefore, focusing on the outcome of interest may often be a better option, although using workflow impact is preferable to relying solely on outcome prediction.

Section 7 — Clinical Adoption & Workflow Impact

7.1 Usability — How easy is it to interpret and use the AI mask compared to GT in clinical workflow?

0 ☐ 1 ☐ 2 ☐ 3 \* 4 ☐ 5 ☐ (0 = much harder, 5 = much easier)

Reason: Although it is smoother and more integrated, making it easier to follow and include features, it still has major failures in certain areas, such as including non-tumoral regions or excluding parts of the tumor.

7.2 Clinical actionability — Would the AI–GT differences change clinical decisions (for example, radiotherapy target, surgery extent, follow-up planning)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 \* (0 = no impact, 5 = strong impact)

Reason: For sure, specialists rely heavily on the ROI, especially for non-invasive procedures like radiotherapy. Even small changes can affect both the size and location of the tumor. Additionally, under-segmentation may cause the model to miss tumors altogether.

7.3 Robustness and variability — How consistent and reliable are AI masks across different cases and observers?

0 ☐ 1 \* 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = very inconsistent, 5 = very consistent)

Reason: Some features remained the same, but for most, the criteria for distinguishing tumoral from non-tumoral areas differed.

7.4 Time-saving potential — How much time could this AI save compared to manual delineation?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 \* 5 ☐ (0 = no time saved, 5 = saves most effort)

Reason: When implemented correctly and supervised to address potential failures, it can be very helpful.

7.5 Trust in clinical use — How confident would you be using AI masks for high-stakes tasks (e.g., radiotherapy planning) without physician refinement?

0 ☐ 1 \* 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no trust, 5 = complete trust)

Reason: the reasons and facts are discussed above

7.6 Confidence maps/interpretability (if available) — Do uncertainty or confidence maps improve your trust in AI outputs?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 \* (0 = no improvement, 5 = very strong improvement)

Reason: as I mentioned ealier this Is very helpful and can improve trust easily.

7.7 Role of AI in clinical practice — In your opinion, AI-based segmentation should:

☐ Replace physicians entirely  
☐ Provide an initial mask for physicians to refine  
\* Assist physicians as a support/check tool

Reason: For now, and likely for the foreseeable future, human knowledge is not extensive enough for algorithms to learn every possible source of noise or anomaly. Even physicians regularly encounter difficult and novel cases. Until human knowledge—or sufficiently diverse datasets—reaches that level, these algorithms should remain as assistants rather than autonomous decision-makers.

**End**

**Comparative Clinical Assessment of Deep Learning Tumor Masks vs. Ground Truth in Lung Cancer**

Model under review: VNet

**Segmentation Results**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Analysis Mode | Model | Dice Score | IoU Score | Hausdorff Distance |
| Test | **VNet** | 0.81 0.08 | 0.69 0.11 | 11.12 3.73 |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

**Prediction Result:**

Radiomics extracted from Ground Truth used for prediction tasks

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Validation | | | | | | Test | | | | | | | Names | | | |
| Acc | Precision | Recall | F1 | AUC | Specificity | Acc | Precision | Recall | F1 | AUC | Specificity | FSA/AEA | | CA | Model | SL/SSL |
| 0.88 | 0.78 | 0.88 | 0.83 | 0.54 | 1 | 0.87 | 0.76 | 0.87 | 0.81 | 0.71 | 1 | Mutual\_Info\_Gain\_Ratio | | ExtraTree | GT | SSL |
| 0.74 | 0.67 | 0.74 | 0.66 | 0.55 | 0.98 | 0.87 | 0.79 | 0.87 | 0.83 | 0.63 | 0.99 | SL |

Radiomics extracted from Vnet used for prediction tasks

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Validation | | | | | | Test | | | | | | Names | | | |
| Acc | Precision | Recall | F1 | AUC | Specificity | Acc | Precision | Recall | F1 | AUC | Specificity | FSA/AEA | CA | Model | SL/SSL |
| 0.88 | 0.78 | 0.88 | 0.83 | 0.65 | 1 | 0.88 | 0.77 | 0.88 | 0.82 | 0.76 | 1 | LASSO | LGBM | VNet | SSL |
| 0.74 | 0.55 | 0.74 | 0.63 | 0.58 | 0.99 | 0.87 | 0.77 | 0.87 | 0.81 | 0.40 | 0.99 | SL |

**Start**

Section 1 — Overall Impression

1.1 Does the AI mask define a tumor volume that is more clinically meaningful than the Ground Truth (GT) for diagnosis and visual interpretation?

0 ☐ 1 ☐ 2 ☐ 3 \* 4 ☐ 5 ☐ (0 = not at all, 5 = very strongly yes)

Reason: It is likely that, to speed up ground truth (GT) segmentation, a high-contrast setting was used, so many of the GTs only include the highest-intensity regions. In contrast, the VNet segmentation appears as a smoother version of the GT, covering a larger peritumoral area. Occasionally, it fails to detect certain features and regions that are important, but sometimes it identifies areas that the GT did not include, which can be significant for diagnosis and interpretation.

Section 2 — Spatial Differences

2.1 Are the AI–GT differences concentrated in the peritumoral zone (≈5–15 mm around the tumor)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 \* (0 = not at all, 5 = strongly concentrated)

Reason: For most of the parts and images this is true.

2.2 How clinically meaningful are the regions where the AI mask differs from GT (including over-segmentation or under-segmentation, such as peritumoral expansion, missed necrotic cores, spiculated margins)? Or, how clinically justified are the AI–GT differences (whether due to additional regions or excluded regions)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 \* 5 ☐

Reason: It depends on the context. For example, if we are predicting a patient’s height, the data would not be meaningful. However, if the outcome of interest is prognosis, the data becomes clinically meaningful and justifiable. For instance, peritumoral areas have been shown to be important for prognosis and for predicting future tumor invasion.

Section 3 — Clinical Interpretation of Differences

3.1 How well does the AI mask preserve internal tumor heterogeneity (for example: solid vs. ground-glass opacity, necrotic vs. viable regions)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 \* 5 ☐

Reason: Both masks preserve heterogeneity, but in the VNet mask, the smoother and relatively larger ROI captures more heterogeneity. Sometimes it highlights regions that appear irrelevant, likely because it includes non-tumoral areas, but at other times, it identifies less dense regions within the tumor that are still important for our analysis.

3.2 Boundary quality: compared to GT, how noisy or fragmented are AI boundaries?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 \* (0 = much less noisy, 5 = much more noisy)

Reason: AI contours are definitely more fragmented and as the result smoother.

Section 4 — Link to Survival Prediction

4.1 Do the AI–GT differences likely capture prognostically relevant tissue that improves survival prediction?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 \* 5 ☐

Reason: Yes, in terms of both heterogeneity and peritumoral areas, this can help in survival prediction. However, the larger ROIs might introduce a bias toward predicting worse survival outcomes.

4.2 How would you rate the effect of the AI’s deviation from GT (larger or smaller volume) on survival prediction?

0 ☐ 1 ☐ 2 ☐ 3 \* 4 ☐ 5 ☐ (0 = strongly harmful, 5 = strongly beneficial)

Reason: I’m on the fence about this. While it may be beneficial in some cases, in others it fails to accurately capture the ROI or produces an unreasonably large ROI.

Section 5 — Trust & Preference (Survival Prediction Focus)

5.1 How likely is it that AI-generated masks are more accurate than GT in some cases (for example, subtle lesions missed by physicians, or over-delineation that reduces noise) in survival prediction tasks?

0 ☐ 1 ☐ 2 \* 3 ☐ 4 ☐ 5 ☐ (0 = very unlikely, 5 = very likely)

Reason: Although it performs better than the GT in capturing peritumoral areas and heterogeneity, if it fails to identify some important tumor regions, it could negatively impact patient care. The size of the active tumor remains one of the most important features, particularly for survival prediction.

5.2 Would you prefer to use the AI mask (rather than GT) as input for survival prediction?

0 ☐ 1 ☐ 2 \* 3 ☐ 4 ☐ 5 ☐ (0 = not at all, 5 = strongly yes)

Reason: I would if there is some explanation of how some very important areas are not included in ROI.

Section 6 — Summaries and Overall Assessment

6.1 Diagnosis-oriented evaluation — How successful is the AI mask in lesion delineation compared to GT?

0 ☐ 1 ☐ 2 ☐ 3 \* 4 ☐ 5 ☐ (0 = complete failure, 5 = highly successful)

Reason: the reasons and facts are discussed above

6.2 Prediction-oriented evaluation — How successful is the AI mask in contributing to survival prediction compared to GT?

0 ☐ 1 ☐ 2 \* 3 ☐ 4 ☐ 5 ☐ (0 = no added value, 5 = very strong added value)

Reason: the reasons and facts are discussed above

6.3 Overall confidence in your assessment

0 ☐ 1 ☐ 2 ☐ 3 \* 4 ☐ 5 ☐ (0 = no confidence, 5 = complete confidence)

Reason: Given that this task ideally requires input from radiologists and oncologists, I am not entirely confident in my assessment.

6.4 Do you think geometric evaluation metrics (e.g., Dice, IoU, Hausdorff distance) are sufficient to establish clinical trust in AI segmentation?

0 ☐ 1 \* 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not sufficient at all, 5 = fully sufficient)

Reason: Actually, the results show the opposite. The key difference is that these AI masks perform better overall. While they can be helpful in certain aspects, such as assessing tumor size, there are still many shortcomings in using these methods.

6.5 What additional types of evaluation do you think are necessary for trustworthy AI segmentation (e.g., radiomics stability, survival prediction, physician review, workflow impact, and others), and why do you think those are effective in building trustworthiness?

Reason: I think radiomics stability still suffers from the same pitfalls as the Dice score. Therefore, focusing on the outcome of interest may often be a better option, although using workflow impact is preferable to relying solely on outcome prediction.

Section 7 — Clinical Adoption & Workflow Impact

7.1 Usability — How easy is it to interpret and use the AI mask compared to GT in clinical workflow?

0 ☐ 1 ☐ 2 \* 3 ☐ 4 ☐ 5 ☐ (0 = much harder, 5 = much easier)

Reason: Although it is smoother and more integrated, making it easier to follow and include features, it still has major failures in certain areas, such as including non-tumoral regions or excluding parts of the tumor.

7.2 Clinical actionability — Would the AI–GT differences change clinical decisions (for example, radiotherapy target, surgery extent, follow-up planning)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 \* (0 = no impact, 5 = strong impact)

Reason: For sure, specialists rely heavily on the ROI, especially for non-invasive procedures like radiotherapy. Even small changes can affect both the size and location of the tumor. Additionally, under-segmentation may cause the model to miss tumors altogether.

7.3 Robustness and variability — How consistent and reliable are AI masks across different cases and observers?

0 ☐ 1 \* 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = very inconsistent, 5 = very consistent)

Reason: Some features remained the same, but for most, the criteria for distinguishing tumoral from non-tumoral areas differed.

7.4 Time-saving potential — How much time could this AI save compared to manual delineation?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 \* 5 ☐ (0 = no time saved, 5 = saves most effort)

Reason: When implemented correctly and supervised to address potential failures, it can be very helpful.

7.5 Trust in clinical use — How confident would you be using AI masks for high-stakes tasks (e.g., radiotherapy planning) without physician refinement?

0 ☐ 1 \* 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no trust, 5 = complete trust)

Reason: the reasons and facts are discussed above

7.6 Confidence maps/interpretability (if available) — Do uncertainty or confidence maps improve your trust in AI outputs?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 \* (0 = no improvement, 5 = very strong improvement)

Reason: as I mentioned ealier this Is very helpful and can improve trust easily.

7.7 Role of AI in clinical practice — In your opinion, AI-based segmentation should:

☐ Replace physicians entirely  
☐ Provide an initial mask for physicians to refine  
\* Assist physicians as a support/check tool

Reason: For now, and likely for the foreseeable future, human knowledge is not extensive enough for algorithms to learn every possible source of noise or anomaly. Even physicians regularly encounter difficult and novel cases. Until human knowledge—or sufficiently diverse datasets—reaches that level, these algorithms should remain as assistants rather than autonomous decision-makers.

**End**

**D5**

**Comparative Clinical Assessment of Deep Learning Tumor Masks vs. Ground Truth in Lung Cancer**

Model under review: 3DAttention

**Segmentation Results**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Analysis Mode | Model | Dice Score | IoU Score | Hausdorff Distance |
| Test |  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
| **3DAttention** | 0.77 0.08 | 0.64 0.1 | 33.61 4.10 |

**Prediction Result:**

Radiomics extracted from Ground Truth used for prediction tasks

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Validation | | | | | | Test | | | | | | | Names | | | |
| Acc | Precision | Recall | F1 | AUC | Specificity | Acc | Precision | Recall | F1 | AUC | Specificity | FSA/AEA | | CA | Model | SL/SSL |
| 0.88 | 0.78 | 0.88 | 0.83 | 0.54 | 1 | 0.87 | 0.76 | 0.87 | 0.81 | 0.71 | 1 | Mutual\_Info\_Gain\_Ratio | | ExtraTree | GT | SSL |
| 0.74 | 0.67 | 0.74 | 0.66 | 0.55 | 0.98 | 0.87 | 0.79 | 0.87 | 0.83 | 0.63 | 0.99 | SL |

Radiomics extracted from DL 3D attention used for prediction tasks

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Validation | | | | | | Test | | | | | | | Names | | | |
| Acc | Precision | Recall | F1 | AUC | Specificity | Acc | Precision | Recall | F1 | AUC | Specificity | FSA/AEA | | CA | Model | SL/SSL |
| 0.79 | 0.80 | 0.79 | 0.80 | 0.69 | 0.87 | 0.81 | 0.83 | 0.81 | 0.82 | 0.76 | 0.88 | Guassian Random Projection | | Voting Classifier | 3DAtt | SSL |
| 0.61 | 0.64 | 0.61 | 0.62 | 0.53 | 0.69 | 0.72 | 0.79 | 0.72 | 0.75 | 0.55 | 0.79 | SL |

**Start**

Section 1 — Overall Impression

1.1 Does the AI mask define a tumor volume that is more clinically meaningful than the Ground Truth (GT) for diagnosis and visual interpretation?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not at all, 5 = very strongly yes)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 2 — Spatial Differences

2.1 Are the AI–GT differences concentrated in the peritumoral zone (≈5–15 mm around the tumor)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not at all, 5 = strongly concentrated)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

2.2 How clinically meaningful are the regions where the AI mask differs from GT (including over-segmentation or under-segmentation, such as peritumoral expansion, missed necrotic cores, spiculated margins)? Or, how clinically justified are the AI–GT differences (whether due to additional regions or excluded regions)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 3 — Clinical Interpretation of Differences

3.1 How well does the AI mask preserve internal tumor heterogeneity (for example: solid vs. ground-glass opacity, necrotic vs. viable regions)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

3.2 Boundary quality: compared to GT, how noisy or fragmented are AI boundaries?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = much less noisy, 5 = much more noisy)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 4 — Link to Survival Prediction

4.1 Do the AI–GT differences likely capture prognostically relevant tissue that improves survival prediction?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

4.2 How would you rate the effect of the AI’s deviation from GT (larger or smaller volume) on survival prediction?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = strongly harmful, 5 = strongly beneficial)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 5 — Trust & Preference (Survival Prediction Focus)

5.1 How likely is it that AI-generated masks are more accurate than GT in some cases (for example, subtle lesions missed by physicians, or over-delineation that reduces noise) in survival prediction tasks?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = very unlikely, 5 = very likely)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

5.2 Would you prefer to use the AI mask (rather than GT) as input for survival prediction?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not at all, 5 = strongly yes)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 6 — Summaries and Overall Assessment

6.1 Diagnosis-oriented evaluation — How successful is the AI mask in lesion delineation compared to GT?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = complete failure, 5 = highly successful)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6.2 Prediction-oriented evaluation — How successful is the AI mask in contributing to survival prediction compared to GT?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no added value, 5 = very strong added value)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6.3 Overall confidence in your assessment

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no confidence, 5 = complete confidence)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6.4 Do you think geometric evaluation metrics (e.g., Dice, IoU, Hausdorff distance) are sufficient to establish clinical trust in AI segmentation?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not sufficient at all, 5 = fully sufficient)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6.5 What additional types of evaluation do you think are necessary for trustworthy AI segmentation (e.g., radiomics stability, survival prediction, physician review, workflow impact, and others), and why do you think those are effective in building trustworthiness?

Survival prediction. Demonstrating alignment with clinical outcomes validates the real-world effectiveness and value of AI segmentation in improving patient care

Section 7 — Clinical Adoption & Workflow Impact

7.1 Usability — How easy is it to interpret and use the AI mask compared to GT in clinical workflow?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = much harder, 5 = much easier)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.2 Clinical actionability — Would the AI–GT differences change clinical decisions (for example, radiotherapy target, surgery extent, follow-up planning)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no impact, 5 = strong impact)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.3 Robustness and variability — How consistent and reliable are AI masks across different cases and observers?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = very inconsistent, 5 = very consistent)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.4 Time-saving potential — How much time could this AI save compared to manual delineation?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no time saved, 5 = saves most effort)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.5 Trust in clinical use — How confident would you be using AI masks for high-stakes tasks (e.g., radiotherapy planning) without physician refinement?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no trust, 5 = complete trust)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.6 Confidence maps/interpretability (if available) — Do uncertainty or confidence maps improve your trust in AI outputs?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no improvement, 5 = very strong improvement)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.7 Role of AI in clinical practice — In your opinion, AI-based segmentation should:

☐ Replace physicians entirely  
☐ Provide an initial mask for physicians to refine  
☐ Assist physicians as a support/check tool

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**End**

**Comparative Clinical Assessment of Deep Learning Tumor Masks vs. Ground Truth in Lung Cancer**

Model under review: 3DMedSam

**Segmentation Results**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Analysis Mode | Model | Dice Score | IoU Score | Hausdorff Distance |
| Test |  |  |  |  |
|  |  |  |  |
|  |  |  |  |
| **3DMedSam** | 0.68 0.19 | 0.57 0.18 | 4.13 1.55 |

**Prediction Result:**

Radiomics extracted from Ground Truth used for prediction tasks

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Validation | | | | | | Test | | | | | | | Names | | | |
| Acc | Precision | Recall | F1 | AUC | Specificity | Acc | Precision | Recall | F1 | AUC | Specificity | FSA/AEA | | CA | Model | SL/SSL |
| 0.88 | 0.78 | 0.88 | 0.83 | 0.54 | 1 | 0.87 | 0.76 | 0.87 | 0.81 | 0.71 | 1 | Mutual\_Info\_Gain\_Ratio | | ExtraTree | GT | SSL |
| 0.74 | 0.67 | 0.74 | 0.66 | 0.55 | 0.98 | 0.87 | 0.79 | 0.87 | 0.83 | 0.63 | 0.99 | SL |

Radiomics extracted from DL 3DmedSam used for prediction tasks

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Validation | | | | | | Test | | | | | | Names | | | |
| Acc | Precision | Recall | F1 | AUC | Specificity | Acc | Precision | Recall | F1 | AUC | Specificity | FSA/AEA | CA | Model | SL/SSL |
| 0.88 | 0.77 | 0.88 | 0.82 | 0.67 | 1 | 0.85 | 0.71 | 0.85 | 0.77 | 0.68 | 1 | ReliefF | GradientBoosting | MedSam | SSL |
| 0.74 | 0.55 | 0.74 | 0.63 | 0.51 | 0.99 | 0.86 | 0.76 | 0.86 | 0.81 | 0.71 | 0.99 | SL |

**Start**

Section 1 — Overall Impression

1.1 Does the AI mask define a tumor volume that is more clinically meaningful than the Ground Truth (GT) for diagnosis and visual interpretation?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not at all, 5 = very strongly yes)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 2 — Spatial Differences

2.1 Are the AI–GT differences concentrated in the peritumoral zone (≈5–15 mm around the tumor)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not at all, 5 = strongly concentrated)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

2.2 How clinically meaningful are the regions where the AI mask differs from GT (including over-segmentation or under-segmentation, such as peritumoral expansion, missed necrotic cores, spiculated margins)? Or, how clinically justified are the AI–GT differences (whether due to additional regions or excluded regions)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 3 — Clinical Interpretation of Differences

3.1 How well does the AI mask preserve internal tumor heterogeneity (for example: solid vs. ground-glass opacity, necrotic vs. viable regions)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

3.2 Boundary quality: compared to GT, how noisy or fragmented are AI boundaries?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = much less noisy, 5 = much more noisy)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 4 — Link to Survival Prediction

4.1 Do the AI–GT differences likely capture prognostically relevant tissue that improves survival prediction?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

4.2 How would you rate the effect of the AI’s deviation from GT (larger or smaller volume) on survival prediction?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = strongly harmful, 5 = strongly beneficial)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 5 — Trust & Preference (Survival Prediction Focus)

5.1 How likely is it that AI-generated masks are more accurate than GT in some cases (for example, subtle lesions missed by physicians, or over-delineation that reduces noise) in survival prediction tasks?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = very unlikely, 5 = very likely)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

5.2 Would you prefer to use the AI mask (rather than GT) as input for survival prediction?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not at all, 5 = strongly yes)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 6 — Summaries and Overall Assessment

6.1 Diagnosis-oriented evaluation — How successful is the AI mask in lesion delineation compared to GT?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = complete failure, 5 = highly successful)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6.2 Prediction-oriented evaluation — How successful is the AI mask in contributing to survival prediction compared to GT?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no added value, 5 = very strong added value)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6.3 Overall confidence in your assessment

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no confidence, 5 = complete confidence)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6.4 Do you think geometric evaluation metrics (e.g., Dice, IoU, Hausdorff distance) are sufficient to establish clinical trust in AI segmentation?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not sufficient at all, 5 = fully sufficient)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6.5 What additional types of evaluation do you think are necessary for trustworthy AI segmentation (e.g., radiomics stability, survival prediction, physician review, workflow impact, and others), and why do you think those are effective in building trustworthiness?

Survival prediction. Demonstrating alignment with clinical outcomes validates the real-world effectiveness and value of AI segmentation in improving patient care

Section 7 — Clinical Adoption & Workflow Impact

7.1 Usability — How easy is it to interpret and use the AI mask compared to GT in clinical workflow?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = much harder, 5 = much easier)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.2 Clinical actionability — Would the AI–GT differences change clinical decisions (for example, radiotherapy target, surgery extent, follow-up planning)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no impact, 5 = strong impact)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.3 Robustness and variability — How consistent and reliable are AI masks across different cases and observers?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = very inconsistent, 5 = very consistent)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.4 Time-saving potential — How much time could this AI save compared to manual delineation?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no time saved, 5 = saves most effort)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.5 Trust in clinical use — How confident would you be using AI masks for high-stakes tasks (e.g., radiotherapy planning) without physician refinement?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no trust, 5 = complete trust)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.6 Confidence maps/interpretability (if available) — Do uncertainty or confidence maps improve your trust in AI outputs?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no improvement, 5 = very strong improvement)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.7 Role of AI in clinical practice — In your opinion, AI-based segmentation should:

☐ Replace physicians entirely  
☐ Provide an initial mask for physicians to refine  
☐ Assist physicians as a support/check tool

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**End**

**Comparative Clinical Assessment of Deep Learning Tumor Masks vs. Ground Truth in Lung Cancer**

Model under review: Reconnet

**Segmentation Results**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Analysis Mode | Model | Dice Score | IoU Score | Hausdorff Distance |
| Test |  |  |  |  |
|  |  |  |  |
| **Reconnet** | 0.79 0.11 | 0.67 0.13 | 11.39 3.96 |
|  |  |  |  |
|  |  |  |  |

**Prediction Result:**

Radiomics extracted from Ground Truth used for prediction tasks

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Validation | | | | | | Test | | | | | | | Names | | | |
| Acc | Precision | Recall | F1 | AUC | Specificity | Acc | Precision | Recall | F1 | AUC | Specificity | FSA/AEA | | CA | Model | SL/SSL |
| 0.88 | 0.78 | 0.88 | 0.83 | 0.54 | 1 | 0.87 | 0.76 | 0.87 | 0.81 | 0.71 | 1 | Mutual\_Info\_Gain\_Ratio | | ExtraTree | GT | SSL |
| 0.74 | 0.67 | 0.74 | 0.66 | 0.55 | 0.98 | 0.87 | 0.79 | 0.87 | 0.83 | 0.63 | 0.99 | SL |

Radiomics extracted from DL Reconnet used for prediction tasks

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Validation | | | | | | Test | | | | | | Names | | | |
| Acc | Precision | Recall | F1 | AUC | Specificity | Acc | Precision | Recall | F1 | AUC | Specificity | FSA/AEA | CA | Model | SL/SSL |
| 0.88 | 0.78 | 0.88 | 0.83 | 0.65 | 1 | 0.86 | 0.75 | 0.86 | 0.80 | 0.71 | 1 | UMAP | AdaBoost | Reconnet | SSL |
| 0.75 | 0.56 | 0.75 | 0.64 | 0.50 | 1 | 0.88 | 0.77 | 0.88 | 0.82 | 0.5 | 1 | SL |

**Start**

Section 1 — Overall Impression

1.1 Does the AI mask define a tumor volume that is more clinically meaningful than the Ground Truth (GT) for diagnosis and visual interpretation?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not at all, 5 = very strongly yes)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 2 — Spatial Differences

2.1 Are the AI–GT differences concentrated in the peritumoral zone (≈5–15 mm around the tumor)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not at all, 5 = strongly concentrated)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

2.2 How clinically meaningful are the regions where the AI mask differs from GT (including over-segmentation or under-segmentation, such as peritumoral expansion, missed necrotic cores, spiculated margins)? Or, how clinically justified are the AI–GT differences (whether due to additional regions or excluded regions)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 3 — Clinical Interpretation of Differences

3.1 How well does the AI mask preserve internal tumor heterogeneity (for example: solid vs. ground-glass opacity, necrotic vs. viable regions)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

3.2 Boundary quality: compared to GT, how noisy or fragmented are AI boundaries?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = much less noisy, 5 = much more noisy)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 4 — Link to Survival Prediction

4.1 Do the AI–GT differences likely capture prognostically relevant tissue that improves survival prediction?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

4.2 How would you rate the effect of the AI’s deviation from GT (larger or smaller volume) on survival prediction?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = strongly harmful, 5 = strongly beneficial)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 5 — Trust & Preference (Survival Prediction Focus)

5.1 How likely is it that AI-generated masks are more accurate than GT in some cases (for example, subtle lesions missed by physicians, or over-delineation that reduces noise) in survival prediction tasks?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = very unlikely, 5 = very likely)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

5.2 Would you prefer to use the AI mask (rather than GT) as input for survival prediction?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not at all, 5 = strongly yes)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 6 — Summaries and Overall Assessment

6.1 Diagnosis-oriented evaluation — How successful is the AI mask in lesion delineation compared to GT?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = complete failure, 5 = highly successful)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6.2 Prediction-oriented evaluation — How successful is the AI mask in contributing to survival prediction compared to GT?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no added value, 5 = very strong added value)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6.3 Overall confidence in your assessment

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no confidence, 5 = complete confidence)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6.4 Do you think geometric evaluation metrics (e.g., Dice, IoU, Hausdorff distance) are sufficient to establish clinical trust in AI segmentation?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not sufficient at all, 5 = fully sufficient)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6.5 What additional types of evaluation do you think are necessary for trustworthy AI segmentation (e.g., radiomics stability, survival prediction, physician review, workflow impact, and others), and why do you think those are effective in building trustworthiness?

Survival prediction. Demonstrating alignment with clinical outcomes validates the real-world effectiveness and value of AI segmentation in improving patient care

Section 7 — Clinical Adoption & Workflow Impact

7.1 Usability — How easy is it to interpret and use the AI mask compared to GT in clinical workflow?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = much harder, 5 = much easier)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.2 Clinical actionability — Would the AI–GT differences change clinical decisions (for example, radiotherapy target, surgery extent, follow-up planning)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no impact, 5 = strong impact)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.3 Robustness and variability — How consistent and reliable are AI masks across different cases and observers?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = very inconsistent, 5 = very consistent)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.4 Time-saving potential — How much time could this AI save compared to manual delineation?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no time saved, 5 = saves most effort)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.5 Trust in clinical use — How confident would you be using AI masks for high-stakes tasks (e.g., radiotherapy planning) without physician refinement?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no trust, 5 = complete trust)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.6 Confidence maps/interpretability (if available) — Do uncertainty or confidence maps improve your trust in AI outputs?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no improvement, 5 = very strong improvement)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.7 Role of AI in clinical practice — In your opinion, AI-based segmentation should:

☐ Replace physicians entirely  
☐ Provide an initial mask for physicians to refine  
☐ Assist physicians as a support/check tool

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**End**

**Comparative Clinical Assessment of Deep Learning Tumor Masks vs. Ground Truth in Lung Cancer**

Model under review: ResUnet

**Segmentation Results**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Analysis Mode | Model | Dice Score | IoU Score | Hausdorff Distance |
| Test |  |  |  |  |
| **ResUNet** | 0.75 0.09 | 0.61 0.1 | 19.91 6.91 |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

**Prediction Result:**

Radiomics extracted from Ground Truth used for prediction tasks

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Validation | | | | | | Test | | | | | | | Names | | | |
| Acc | Precision | Recall | F1 | AUC | Specificity | Acc | Precision | Recall | F1 | AUC | Specificity | FSA/AEA | | CA | Model | SL/SSL |
| 0.88 | 0.78 | 0.88 | 0.83 | 0.54 | 1 | 0.87 | 0.76 | 0.87 | 0.81 | 0.71 | 1 | Mutual\_Info\_Gain\_Ratio | | ExtraTree | GT | SSL |
| 0.74 | 0.67 | 0.74 | 0.66 | 0.55 | 0.98 | 0.87 | 0.79 | 0.87 | 0.83 | 0.63 | 0.99 | SL |

Radiomics extracted from DL ResUNet used for prediction tasks

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Validation | | | | | | Test | | | | | | | Names | | | |
| Acc | Precision | Recall | F1 | AUC | Specificity | Acc | Precision | Recall | F1 | AUC | Specificity | FSA/AEA | | CA | Model | SL/SSL |
| 0.88 | 0.78 | 0.88 | 0.83 | 0.30 | 1 | 0.84 | 0.71 | 0.84 | 0.77 | 0.67 | 1 | Mutual Information | | MLP | ResUNet | SSL |
| 0.75 | 0.56 | 0.75 | 0.64 | 0.56 | 1 | 0.83 | 0.69 | 0.83 | 0.75 | 0.68 | 1 | SL |

**Start**

Section 1 — Overall Impression

1.1 Does the AI mask define a tumor volume that is more clinically meaningful than the Ground Truth (GT) for diagnosis and visual interpretation?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not at all, 5 = very strongly yes)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 2 — Spatial Differences

2.1 Are the AI–GT differences concentrated in the peritumoral zone (≈5–15 mm around the tumor)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not at all, 5 = strongly concentrated)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

2.2 How clinically meaningful are the regions where the AI mask differs from GT (including over-segmentation or under-segmentation, such as peritumoral expansion, missed necrotic cores, spiculated margins)? Or, how clinically justified are the AI–GT differences (whether due to additional regions or excluded regions)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 3 — Clinical Interpretation of Differences

3.1 How well does the AI mask preserve internal tumor heterogeneity (for example: solid vs. ground-glass opacity, necrotic vs. viable regions)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

3.2 Boundary quality: compared to GT, how noisy or fragmented are AI boundaries?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = much less noisy, 5 = much more noisy)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 4 — Link to Survival Prediction

4.1 Do the AI–GT differences likely capture prognostically relevant tissue that improves survival prediction?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

4.2 How would you rate the effect of the AI’s deviation from GT (larger or smaller volume) on survival prediction?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = strongly harmful, 5 = strongly beneficial)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 5 — Trust & Preference (Survival Prediction Focus)

5.1 How likely is it that AI-generated masks are more accurate than GT in some cases (for example, subtle lesions missed by physicians, or over-delineation that reduces noise) in survival prediction tasks?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = very unlikely, 5 = very likely)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

5.2 Would you prefer to use the AI mask (rather than GT) as input for survival prediction?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not at all, 5 = strongly yes)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 6 — Summaries and Overall Assessment

6.1 Diagnosis-oriented evaluation — How successful is the AI mask in lesion delineation compared to GT?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = complete failure, 5 = highly successful)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6.2 Prediction-oriented evaluation — How successful is the AI mask in contributing to survival prediction compared to GT?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no added value, 5 = very strong added value)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6.3 Overall confidence in your assessment

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no confidence, 5 = complete confidence)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6.4 Do you think geometric evaluation metrics (e.g., Dice, IoU, Hausdorff distance) are sufficient to establish clinical trust in AI segmentation?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not sufficient at all, 5 = fully sufficient)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6.5 What additional types of evaluation do you think are necessary for trustworthy AI segmentation (e.g., radiomics stability, survival prediction, physician review, workflow impact, and others), and why do you think those are effective in building trustworthiness?

Survival prediction. Demonstrating alignment with clinical outcomes validates the real-world effectiveness and value of AI segmentation in improving patient care

Section 7 — Clinical Adoption & Workflow Impact

7.1 Usability — How easy is it to interpret and use the AI mask compared to GT in clinical workflow?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = much harder, 5 = much easier)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.2 Clinical actionability — Would the AI–GT differences change clinical decisions (for example, radiotherapy target, surgery extent, follow-up planning)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no impact, 5 = strong impact)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.3 Robustness and variability — How consistent and reliable are AI masks across different cases and observers?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = very inconsistent, 5 = very consistent)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.4 Time-saving potential — How much time could this AI save compared to manual delineation?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no time saved, 5 = saves most effort)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.5 Trust in clinical use — How confident would you be using AI masks for high-stakes tasks (e.g., radiotherapy planning) without physician refinement?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no trust, 5 = complete trust)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.6 Confidence maps/interpretability (if available) — Do uncertainty or confidence maps improve your trust in AI outputs?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no improvement, 5 = very strong improvement)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.7 Role of AI in clinical practice — In your opinion, AI-based segmentation should:

☐ Replace physicians entirely  
☐ Provide an initial mask for physicians to refine  
☐ Assist physicians as a support/check tool

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**End**

**Comparative Clinical Assessment of Deep Learning Tumor Masks vs. Ground Truth in Lung Cancer**

Model under review: VNet

**Segmentation Results**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Analysis Mode | Model | Dice Score | IoU Score | Hausdorff Distance |
| Test | **VNet** | 0.83 0.07 | 0.71 0.09 | 10.04 3.73 |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

**Prediction Result:**

Radiomics extracted from Ground Truth used for prediction tasks

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Validation | | | | | | Test | | | | | | | Names | | | |
| Acc | Precision | Recall | F1 | AUC | Specificity | Acc | Precision | Recall | F1 | AUC | Specificity | FSA/AEA | | CA | Model | SL/SSL |
| 0.88 | 0.78 | 0.88 | 0.83 | 0.54 | 1 | 0.87 | 0.76 | 0.87 | 0.81 | 0.71 | 1 | Mutual\_Info\_Gain\_Ratio | | ExtraTree | GT | SSL |
| 0.74 | 0.67 | 0.74 | 0.66 | 0.55 | 0.98 | 0.87 | 0.79 | 0.87 | 0.83 | 0.63 | 0.99 | SL |

Radiomics extracted from Vnet used for prediction tasks

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Validation | | | | | | Test | | | | | | Names | | | |
| Acc | Precision | Recall | F1 | AUC | Specificity | Acc | Precision | Recall | F1 | AUC | Specificity | FSA/AEA | CA | Model | SL/SSL |
| 0.88 | 0.78 | 0.88 | 0.83 | 0.65 | 1 | 0.88 | 0.77 | 0.88 | 0.82 | 0.76 | 1 | LASSO | LGBM | VNet | SSL |
| 0.74 | 0.55 | 0.74 | 0.63 | 0.58 | 0.99 | 0.87 | 0.77 | 0.87 | 0.81 | 0.40 | 0.99 | SL |

**Start**

Section 1 — Overall Impression

1.1 Does the AI mask define a tumor volume that is more clinically meaningful than the Ground Truth (GT) for diagnosis and visual interpretation?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not at all, 5 = very strongly yes)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 2 — Spatial Differences

2.1 Are the AI–GT differences concentrated in the peritumoral zone (≈5–15 mm around the tumor)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not at all, 5 = strongly concentrated)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

2.2 How clinically meaningful are the regions where the AI mask differs from GT (including over-segmentation or under-segmentation, such as peritumoral expansion, missed necrotic cores, spiculated margins)? Or, how clinically justified are the AI–GT differences (whether due to additional regions or excluded regions)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 3 — Clinical Interpretation of Differences

3.1 How well does the AI mask preserve internal tumor heterogeneity (for example: solid vs. ground-glass opacity, necrotic vs. viable regions)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

3.2 Boundary quality: compared to GT, how noisy or fragmented are AI boundaries?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = much less noisy, 5 = much more noisy)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 4 — Link to Survival Prediction

4.1 Do the AI–GT differences likely capture prognostically relevant tissue that improves survival prediction?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

4.2 How would you rate the effect of the AI’s deviation from GT (larger or smaller volume) on survival prediction?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = strongly harmful, 5 = strongly beneficial)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 5 — Trust & Preference (Survival Prediction Focus)

5.1 How likely is it that AI-generated masks are more accurate than GT in some cases (for example, subtle lesions missed by physicians, or over-delineation that reduces noise) in survival prediction tasks?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = very unlikely, 5 = very likely)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

5.2 Would you prefer to use the AI mask (rather than GT) as input for survival prediction?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not at all, 5 = strongly yes)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 6 — Summaries and Overall Assessment

6.1 Diagnosis-oriented evaluation — How successful is the AI mask in lesion delineation compared to GT?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = complete failure, 5 = highly successful)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6.2 Prediction-oriented evaluation — How successful is the AI mask in contributing to survival prediction compared to GT?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no added value, 5 = very strong added value)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6.3 Overall confidence in your assessment

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no confidence, 5 = complete confidence)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6.4 Do you think geometric evaluation metrics (e.g., Dice, IoU, Hausdorff distance) are sufficient to establish clinical trust in AI segmentation?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not sufficient at all, 5 = fully sufficient)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6.5 What additional types of evaluation do you think are necessary for trustworthy AI segmentation (e.g., radiomics stability, survival prediction, physician review, workflow impact, and others), and why do you think those are effective in building trustworthiness?

Survival prediction. Demonstrating alignment with clinical outcomes validates the real-world effectiveness and value of AI segmentation in improving patient care

Section 7 — Clinical Adoption & Workflow Impact

7.1 Usability — How easy is it to interpret and use the AI mask compared to GT in clinical workflow?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = much harder, 5 = much easier)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.2 Clinical actionability — Would the AI–GT differences change clinical decisions (for example, radiotherapy target, surgery extent, follow-up planning)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no impact, 5 = strong impact)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.3 Robustness and variability — How consistent and reliable are AI masks across different cases and observers?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = very inconsistent, 5 = very consistent)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.4 Time-saving potential — How much time could this AI save compared to manual delineation?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no time saved, 5 = saves most effort)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.5 Trust in clinical use — How confident would you be using AI masks for high-stakes tasks (e.g., radiotherapy planning) without physician refinement?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no trust, 5 = complete trust)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.6 Confidence maps/interpretability (if available) — Do uncertainty or confidence maps improve your trust in AI outputs?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no improvement, 5 = very strong improvement)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.7 Role of AI in clinical practice — In your opinion, AI-based segmentation should:

☐ Replace physicians entirely  
☐ Provide an initial mask for physicians to refine  
☐ Assist physicians as a support/check tool

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**End**

**D6**

**Comparative Clinical Assessment of Deep Learning Tumor Masks vs. Ground Truth in Lung Cancer**

Model under review: 3DAttention

**Segmentation Results**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Analysis Mode | Model | Dice Score | IoU Score | Hausdorff Distance |
| Test |  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
| **3DAttention** | 0.83 0.04 | 0.72 0.04 | 12.12 3.73 |

**Prediction Result:**

Radiomics extracted from Ground Truth used for prediction tasks

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Validation | | | | | | Test | | | | | | | Names | | | |
| Acc | Precision | Recall | F1 | AUC | Specificity | Acc | Precision | Recall | F1 | AUC | Specificity | FSA/AEA | | CA | Model | SL/SSL |
| 0.88 | 0.78 | 0.88 | 0.83 | 0.54 | 1 | 0.87 | 0.76 | 0.87 | 0.81 | 0.71 | 1 | Mutual\_Info\_Gain\_Ratio | | ExtraTree | GT | SSL |
| 0.74 | 0.67 | 0.74 | 0.66 | 0.55 | 0.98 | 0.87 | 0.79 | 0.87 | 0.83 | 0.63 | 0.99 | SL |

Radiomics extracted from DL 3D attention used for prediction tasks

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Validation | | | | | | Test | | | | | | | Names | | | |
| Acc | Precision | Recall | F1 | AUC | Specificity | Acc | Precision | Recall | F1 | AUC | Specificity | FSA/AEA | | CA | Model | SL/SSL |
| 0.79 | 0.80 | 0.79 | 0.80 | 0.69 | 0.87 | 0.81 | 0.83 | 0.81 | 0.82 | 0.76 | 0.88 | Guassian Random Projection | | Voting Classifier | 3DAtt | SSL |
| 0.61 | 0.64 | 0.61 | 0.62 | 0.53 | 0.69 | 0.72 | 0.79 | 0.72 | 0.75 | 0.55 | 0.79 | SL |

**Start**

Section 1 — Overall Impression

1.1 Does the AI mask define a tumor volume that is more clinically meaningful than the Ground Truth (GT) for diagnosis and visual interpretation?

0 ☐ 1 ☐ 2 x 3 ☐ 4 ☐ 5 ☐ (0 = not at all, 5 = very strongly yes)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 2 — Spatial Differences

2.1 Are the AI–GT differences concentrated in the peritumoral zone (≈5–15 mm around the tumor)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 x 5 ☐ (0 = not at all, 5 = strongly concentrated)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

2.2 How clinically meaningful are the regions where the AI mask differs from GT (including over-segmentation or under-segmentation, such as peritumoral expansion, missed necrotic cores, spiculated margins)? Or, how clinically justified are the AI–GT differences (whether due to additional regions or excluded regions)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 x 5 ☐

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 3 — Clinical Interpretation of Differences

3.1 How well does the AI mask preserve internal tumor heterogeneity (for example: solid vs. ground-glass opacity, necrotic vs. viable regions)?

0 ☐ 1 x 2 ☐ 3 ☐ 4 ☐ 5 ☐

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

3.2 Boundary quality: compared to GT, how noisy or fragmented are AI boundaries?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 x (0 = much less noisy, 5 = much more noisy)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 4 — Link to Survival Prediction

4.1 Do the AI–GT differences likely capture prognostically relevant tissue that improves survival prediction?

0 x 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

4.2 How would you rate the effect of the AI’s deviation from GT (larger or smaller volume) on survival prediction?

0 ☐ 1 ☐ 2 x 3 ☐ 4 ☐ 5 ☐ (0 = strongly harmful, 5 = strongly beneficial)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 5 — Trust & Preference (Survival Prediction Focus)

5.1 How likely is it that AI-generated masks are more accurate than GT in some cases (for example, subtle lesions missed by physicians, or over-delineation that reduces noise) in survival prediction tasks?

0 ☐ 1 x 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = very unlikely, 5 = very likely)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

5.2 Would you prefer to use the AI mask (rather than GT) as input for survival prediction?

0 ☐ 1 x 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not at all, 5 = strongly yes)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 6 — Summaries and Overall Assessment

6.1 Diagnosis-oriented evaluation — How successful is the AI mask in lesion delineation compared to GT?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 x 5 ☐ (0 = complete failure, 5 = highly successful)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6.2 Prediction-oriented evaluation — How successful is the AI mask in contributing to survival prediction compared to GT?

0 x 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no added value, 5 = very strong added value)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6.3 Overall confidence in your assessment

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 x 5 ☐ (0 = no confidence, 5 = complete confidence)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6.4 Do you think geometric evaluation metrics (e.g., Dice, IoU, Hausdorff distance) are sufficient to establish clinical trust in AI segmentation?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 x 5 ☐ (0 = not sufficient at all, 5 = fully sufficient)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6.5 What additional types of evaluation do you think are necessary for trustworthy AI segmentation (e.g., radiomics stability, survival prediction, physician review, workflow impact, and others), and why do you think those are effective in building trustworthiness?

Reason: \_evaluation of random noises in the segments and fully registered segmentations with the CT as well as an easy-to-use workflow\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 7 — Clinical Adoption & Workflow Impact

7.1 Usability — How easy is it to interpret and use the AI mask compared to GT in clinical workflow?

0 ☐ 1 ☐ 2 x 3 ☐ 4 ☐ 5 ☐ (0 = much harder, 5 = much easier)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.2 Clinical actionability — Would the AI–GT differences change clinical decisions (for example, radiotherapy target, surgery extent, follow-up planning)?

0 ☐ 1 ☐ 2 ☐ 3 x 4 ☐ 5 ☐ (0 = no impact, 5 = strong impact)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.3 Robustness and variability — How consistent and reliable are AI masks across different cases and observers?

0 ☐ 1 ☐ 2 x3 ☐ 4 ☐ 5 ☐ (0 = very inconsistent, 5 = very consistent)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.4 Time-saving potential — How much time could this AI save compared to manual delineation?

0 ☐ 1 ☐ 2 ☐ 3 x 4 ☐ 5 ☐ (0 = no time saved, 5 = saves most effort)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.5 Trust in clinical use — How confident would you be using AI masks for high-stakes tasks (e.g., radiotherapy planning) without physician refinement?

0 ☐1 x 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no trust, 5 = complete trust)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.6 Confidence maps/interpretability (if available) — Do uncertainty or confidence maps improve your trust in AI outputs?

0 ☐ 1 ☐ 2 ☐ 3 x 4 ☐ 5 ☐ (0 = no improvement, 5 = very strong improvement)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.7 Role of AI in clinical practice — In your opinion, AI-based segmentation should:

☐ Replace physicians entirely  
☐ Provide an initial mask for physicians to refine  
x Assist physicians as a support/check tool

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**End**

**Comparative Clinical Assessment of Deep Learning Tumor Masks vs. Ground Truth in Lung Cancer**

Model under review: 3DMedSam

**Segmentation Results**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Analysis Mode | Model | Dice Score | IoU Score | Hausdorff Distance |
| Test |  |  |  |  |
|  |  |  |  |
|  |  |  |  |
| **3DMedSam** | 0.68 0.19 | 0.52 0.18 | 4.26 1.55 |

**Prediction Result:**

Radiomics extracted from Ground Truth used for prediction tasks

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Validation | | | | | | Test | | | | | | | Names | | | |
| Acc | Precision | Recall | F1 | AUC | Specificity | Acc | Precision | Recall | F1 | AUC | Specificity | FSA/AEA | | CA | Model | SL/SSL |
| 0.88 | 0.78 | 0.88 | 0.83 | 0.54 | 1 | 0.87 | 0.76 | 0.87 | 0.81 | 0.71 | 1 | Mutual\_Info\_Gain\_Ratio | | ExtraTree | GT | SSL |
| 0.74 | 0.67 | 0.74 | 0.66 | 0.55 | 0.98 | 0.87 | 0.79 | 0.87 | 0.83 | 0.63 | 0.99 | SL |

Radiomics extracted from DL 3DmedSam used for prediction tasks

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Validation | | | | | | Test | | | | | | Names | | | |
| Acc | Precision | Recall | F1 | AUC | Specificity | Acc | Precision | Recall | F1 | AUC | Specificity | FSA/AEA | CA | Model | SL/SSL |
| 0.88 | 0.77 | 0.88 | 0.82 | 0.67 | 1 | 0.85 | 0.71 | 0.85 | 0.77 | 0.68 | 1 | ReliefF | GradientBoosting | MedSam | SSL |
| 0.74 | 0.55 | 0.74 | 0.63 | 0.51 | 0.99 | 0.86 | 0.76 | 0.86 | 0.81 | 0.71 | 0.99 | SL |

**Start**

Section 1 — Overall Impression

1.1 Does the AI mask define a tumor volume that is more clinically meaningful than the Ground Truth (GT) for diagnosis and visual interpretation?

0 ☐ 1 ☐ 2 ☐ 3 x 4 ☐ 5 ☐ (0 = not at all, 5 = very strongly yes)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 2 — Spatial Differences

2.1 Are the AI–GT differences concentrated in the peritumoral zone (≈5–15 mm around the tumor)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 x (0 = not at all, 5 = strongly concentrated)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

2.2 How clinically meaningful are the regions where the AI mask differs from GT (including over-segmentation or under-segmentation, such as peritumoral expansion, missed necrotic cores, spiculated margins)? Or, how clinically justified are the AI–GT differences (whether due to additional regions or excluded regions)?

0 ☐ 1 x 2 ☐ 3 ☐ 4 ☐ 5 ☐

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 3 — Clinical Interpretation of Differences

3.1 How well does the AI mask preserve internal tumor heterogeneity (for example: solid vs. ground-glass opacity, necrotic vs. viable regions)?

0 ☐ 1 ☐ 2 ☐ 3 x 4 ☐ 5 ☐

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

3.2 Boundary quality: compared to GT, how noisy or fragmented are AI boundaries?

0 ☐ 1 ☐ 2 ☐ 3 x 4 ☐ 5 ☐ (0 = much less noisy, 5 = much more noisy)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 4 — Link to Survival Prediction

4.1 Do the AI–GT differences likely capture prognostically relevant tissue that improves survival prediction?

0 x1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

4.2 How would you rate the effect of the AI’s deviation from GT (larger or smaller volume) on survival prediction?

0 ☐ 1 ☐ 2 ☐ 3 x 4 ☐ 5 ☐ (0 = strongly harmful, 5 = strongly beneficial)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 5 — Trust & Preference (Survival Prediction Focus)

5.1 How likely is it that AI-generated masks are more accurate than GT in some cases (for example, subtle lesions missed by physicians, or over-delineation that reduces noise) in survival prediction tasks?

0 ☐ 1 ☐ 2 ☐ 3 x 4 ☐ 5 ☐ (0 = very unlikely, 5 = very likely)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

5.2 Would you prefer to use the AI mask (rather than GT) as input for survival prediction?

0 ☐ 1 ☐ 2 x 3 ☐ 4 ☐ 5 ☐ (0 = not at all, 5 = strongly yes)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 6 — Summaries and Overall Assessment

6.1 Diagnosis-oriented evaluation — How successful is the AI mask in lesion delineation compared to GT?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 x (0 = complete failure, 5 = highly successful)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6.2 Prediction-oriented evaluation — How successful is the AI mask in contributing to survival prediction compared to GT?

0 ☐ 1 ☐ 2 x 3 ☐ 4 ☐ 5 ☐ (0 = no added value, 5 = very strong added value)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6.3 Overall confidence in your assessment

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 x 5 ☐ (0 = no confidence, 5 = complete confidence)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6.4 Do you think geometric evaluation metrics (e.g., Dice, IoU, Hausdorff distance) are sufficient to establish clinical trust in AI segmentation?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 x 5 ☐ (0 = not sufficient at all, 5 = fully sufficient)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6.5 What additional types of evaluation do you think are necessary for trustworthy AI segmentation (e.g., radiomics stability, survival prediction, physician review, workflow impact, and others), and why do you think those are effective in building trustworthiness?

Reason: \_evaluation of random noises in the segments and fully registered segmentations with the CT as well as an easy-to-use workflow\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 7 — Clinical Adoption & Workflow Impact

7.1 Usability — How easy is it to interpret and use the AI mask compared to GT in clinical workflow?

0 ☐ 1 ☐ 2 x 3 ☐ 4 ☐ 5 ☐ (0 = much harder, 5 = much easier)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.2 Clinical actionability — Would the AI–GT differences change clinical decisions (for example, radiotherapy target, surgery extent, follow-up planning)?

0 ☐ 1 ☐ 2 ☐ 3 x 4 ☐ 5 ☐ (0 = no impact, 5 = strong impact)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.3 Robustness and variability — How consistent and reliable are AI masks across different cases and observers?

0 ☐ 1 ☐ 2 x3 ☐ 4 ☐ 5 ☐ (0 = very inconsistent, 5 = very consistent)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.4 Time-saving potential — How much time could this AI save compared to manual delineation?

0 ☐ 1 ☐ 2 ☐ 3 x 4 ☐ 5 ☐ (0 = no time saved, 5 = saves most effort)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.5 Trust in clinical use — How confident would you be using AI masks for high-stakes tasks (e.g., radiotherapy planning) without physician refinement?

0 ☐1 x 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no trust, 5 = complete trust)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.6 Confidence maps/interpretability (if available) — Do uncertainty or confidence maps improve your trust in AI outputs?

0 ☐ 1 ☐ 2 ☐ 3 x 4 ☐ 5 ☐ (0 = no improvement, 5 = very strong improvement)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.7 Role of AI in clinical practice — In your opinion, AI-based segmentation should:

☐ Replace physicians entirely  
☐ Provide an initial mask for physicians to refine  
x Assist physicians as a support/check tool

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**End**

**Comparative Clinical Assessment of Deep Learning Tumor Masks vs. Ground Truth in Lung Cancer**

Model under review: Reconnet

**Segmentation Results**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Analysis Mode | Model | Dice Score | IoU Score | Hausdorff Distance |
| Test |  |  |  |  |
|  |  |  |  |
| **Reconnet** | 0.80 0.09 | 0.68 0.13 | 11.30 3.96 |
|  |  |  |  |
|  |  |  |  |

**Prediction Result:**

Radiomics extracted from Ground Truth used for prediction tasks

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Validation | | | | | | Test | | | | | | | Names | | | |
| Acc | Precision | Recall | F1 | AUC | Specificity | Acc | Precision | Recall | F1 | AUC | Specificity | FSA/AEA | | CA | Model | SL/SSL |
| 0.88 | 0.78 | 0.88 | 0.83 | 0.54 | 1 | 0.87 | 0.76 | 0.87 | 0.81 | 0.71 | 1 | Mutual\_Info\_Gain\_Ratio | | ExtraTree | GT | SSL |
| 0.74 | 0.67 | 0.74 | 0.66 | 0.55 | 0.98 | 0.87 | 0.79 | 0.87 | 0.83 | 0.63 | 0.99 | SL |

Radiomics extracted from DL Reconnet used for prediction tasks

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Validation | | | | | | Test | | | | | | Names | | | |
| Acc | Precision | Recall | F1 | AUC | Specificity | Acc | Precision | Recall | F1 | AUC | Specificity | FSA/AEA | CA | Model | SL/SSL |
| 0.88 | 0.78 | 0.88 | 0.83 | 0.65 | 1 | 0.86 | 0.75 | 0.86 | 0.80 | 0.71 | 1 | UMAP | AdaBoost | Reconnet | SSL |
| 0.75 | 0.56 | 0.75 | 0.64 | 0.50 | 1 | 0.88 | 0.77 | 0.88 | 0.82 | 0.5 | 1 | SL |

**Start**

Section 1 — Overall Impression

1.1 Does the AI mask define a tumor volume that is more clinically meaningful than the Ground Truth (GT) for diagnosis and visual interpretation?

0 x1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not at all, 5 = very strongly yes)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 2 — Spatial Differences

2.1 Are the AI–GT differences concentrated in the peritumoral zone (≈5–15 mm around the tumor)?

0 ☐ 1 x 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not at all, 5 = strongly concentrated)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

2.2 How clinically meaningful are the regions where the AI mask differs from GT (including over-segmentation or under-segmentation, such as peritumoral expansion, missed necrotic cores, spiculated margins)? Or, how clinically justified are the AI–GT differences (whether due to additional regions or excluded regions)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 x

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 3 — Clinical Interpretation of Differences

3.1 How well does the AI mask preserve internal tumor heterogeneity (for example: solid vs. ground-glass opacity, necrotic vs. viable regions)?

0 ☐ 1 x 2 ☐ 3 ☐ 4 ☐ 5 ☐

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

3.2 Boundary quality: compared to GT, how noisy or fragmented are AI boundaries?

0 ☐ 1 ☐ 2 ☐ 3 x 4 ☐ 5 ☐ (0 = much less noisy, 5 = much more noisy)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 4 — Link to Survival Prediction

4.1 Do the AI–GT differences likely capture prognostically relevant tissue that improves survival prediction?

0 ☐ 1 x 2 ☐ 3 ☐ 4 ☐ 5 ☐

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

4.2 How would you rate the effect of the AI’s deviation from GT (larger or smaller volume) on survival prediction?

0 ☐ 1 x 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = strongly harmful, 5 = strongly beneficial)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 5 — Trust & Preference (Survival Prediction Focus)

5.1 How likely is it that AI-generated masks are more accurate than GT in some cases (for example, subtle lesions missed by physicians, or over-delineation that reduces noise) in survival prediction tasks?

0 ☐ 1 x 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = very unlikely, 5 = very likely)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

5.2 Would you prefer to use the AI mask (rather than GT) as input for survival prediction?

0 x 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not at all, 5 = strongly yes)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 6 — Summaries and Overall Assessment

6.1 Diagnosis-oriented evaluation — How successful is the AI mask in lesion delineation compared to GT?

0 ☐ 1 ☐ 2 x 3 ☐ 4 ☐ 5 ☐ (0 = complete failure, 5 = highly successful)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6.2 Prediction-oriented evaluation — How successful is the AI mask in contributing to survival prediction compared to GT?

0 x 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no added value, 5 = very strong added value)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6.3 Overall confidence in your assessment

0 ☐ 1 ☐ 2 ☐ 3 x 4 ☐ 5 ☐ (0 = no confidence, 5 = complete confidence)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6.4 Do you think geometric evaluation metrics (e.g., Dice, IoU, Hausdorff distance) are sufficient to establish clinical trust in AI segmentation?

0 ☐ 1 ☐ 2 ☐ 3 x 4 ☐ 5 ☐ (0 = not sufficient at all, 5 = fully sufficient)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6.5 What additional types of evaluation do you think are necessary for trustworthy AI segmentation (e.g., radiomics stability, survival prediction, physician review, workflow impact, and others), and why do you think those are effective in building trustworthiness?

Reason: \_evaluation of random noises in the segments and fully registered segmentations with the CT as well as an easy-to-use workflow\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 7 — Clinical Adoption & Workflow Impact

7.1 Usability — How easy is it to interpret and use the AI mask compared to GT in clinical workflow?

0 ☐ 1 ☐ 2 x 3 ☐ 4 ☐ 5 ☐ (0 = much harder, 5 = much easier)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.2 Clinical actionability — Would the AI–GT differences change clinical decisions (for example, radiotherapy target, surgery extent, follow-up planning)?

0 ☐ 1 ☐ 2 ☐ 3 x 4 ☐ 5 ☐ (0 = no impact, 5 = strong impact)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.3 Robustness and variability — How consistent and reliable are AI masks across different cases and observers?

0 ☐ 1 ☐ 2 x3 ☐ 4 ☐ 5 ☐ (0 = very inconsistent, 5 = very consistent)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.4 Time-saving potential — How much time could this AI save compared to manual delineation?

0 ☐ 1 ☐ 2 ☐ 3 x 4 ☐ 5 ☐ (0 = no time saved, 5 = saves most effort)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.5 Trust in clinical use — How confident would you be using AI masks for high-stakes tasks (e.g., radiotherapy planning) without physician refinement?

0 ☐1 x 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no trust, 5 = complete trust)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.6 Confidence maps/interpretability (if available) — Do uncertainty or confidence maps improve your trust in AI outputs?

0 ☐ 1 ☐ 2 ☐ 3 x 4 ☐ 5 ☐ (0 = no improvement, 5 = very strong improvement)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.7 Role of AI in clinical practice — In your opinion, AI-based segmentation should:

☐ Replace physicians entirely  
☐ Provide an initial mask for physicians to refine  
x Assist physicians as a support/check tool

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**End**

**Comparative Clinical Assessment of Deep Learning Tumor Masks vs. Ground Truth in Lung Cancer**

Model under review: ResUnet

**Segmentation Results**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Analysis Mode | Model | Dice Score | IoU Score | Hausdorff Distance |
| Test |  |  |  |  |
| **ResUNet** | 75 0.08 | 0.61 0.10 | 17.95 6.91 |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

**Prediction Result:**

Radiomics extracted from Ground Truth used for prediction tasks

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Validation | | | | | | Test | | | | | | | Names | | | |
| Acc | Precision | Recall | F1 | AUC | Specificity | Acc | Precision | Recall | F1 | AUC | Specificity | FSA/AEA | | CA | Model | SL/SSL |
| 0.88 | 0.78 | 0.88 | 0.83 | 0.54 | 1 | 0.87 | 0.76 | 0.87 | 0.81 | 0.71 | 1 | Mutual\_Info\_Gain\_Ratio | | ExtraTree | GT | SSL |
| 0.74 | 0.67 | 0.74 | 0.66 | 0.55 | 0.98 | 0.87 | 0.79 | 0.87 | 0.83 | 0.63 | 0.99 | SL |

Radiomics extracted from DL ResUNet used for prediction tasks

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Validation | | | | | | Test | | | | | | | Names | | | |
| Acc | Precision | Recall | F1 | AUC | Specificity | Acc | Precision | Recall | F1 | AUC | Specificity | FSA/AEA | | CA | Model | SL/SSL |
| 0.88 | 0.78 | 0.88 | 0.83 | 0.30 | 1 | 0.84 | 0.71 | 0.84 | 0.77 | 0.67 | 1 | Mutual Information | | MLP | ResUNet | SSL |
| 0.75 | 0.56 | 0.75 | 0.64 | 0.56 | 1 | 0.83 | 0.69 | 0.83 | 0.75 | 0.68 | 1 | SL |

**Start**

Section 1 — Overall Impression

1.1 Does the AI mask define a tumor volume that is more clinically meaningful than the Ground Truth (GT) for diagnosis and visual interpretation?

0 ☐ 1 X 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = not at all, 5 = very strongly yes)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 2 — Spatial Differences

2.1 Are the AI–GT differences concentrated in the peritumoral zone (≈5–15 mm around the tumor)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 X 5 ☐ (0 = not at all, 5 = strongly concentrated)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

2.2 How clinically meaningful are the regions where the AI mask differs from GT (including over-segmentation or under-segmentation, such as peritumoral expansion, missed necrotic cores, spiculated margins)? Or, how clinically justified are the AI–GT differences (whether due to additional regions or excluded regions)?

0 ☐ 1 ☐ 2 ☐ 3 X 4 ☐ 5 ☐

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 3 — Clinical Interpretation of Differences

3.1 How well does the AI mask preserve internal tumor heterogeneity (for example: solid vs. ground-glass opacity, necrotic vs. viable regions)?

0 ☐ 1 ☐ 2 ☐ 3 X 4 ☐ 5 ☐

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

3.2 Boundary quality: compared to GT, how noisy or fragmented are AI boundaries?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 X 5 ☐ (0 = much less noisy, 5 = much more noisy)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 4 — Link to Survival Prediction

4.1 Do the AI–GT differences likely capture prognostically relevant tissue that improves survival prediction?

0 ☐ 1 X 2 ☐ 3 ☐ 4 ☐ 5 ☐

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

4.2 How would you rate the effect of the AI’s deviation from GT (larger or smaller volume) on survival prediction?

0 ☐ 1 ☐ 2 X 3 ☐ 4 ☐ 5 ☐ (0 = strongly harmful, 5 = strongly beneficial)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 5 — Trust & Preference (Survival Prediction Focus)

5.1 How likely is it that AI-generated masks are more accurate than GT in some cases (for example, subtle lesions missed by physicians, or over-delineation that reduces noise) in survival prediction tasks?

0 ☐ 1 ☐ 2 X 3 ☐ 4 ☐ 5 ☐ (0 = very unlikely, 5 = very likely)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

5.2 Would you prefer to use the AI mask (rather than GT) as input for survival prediction?

0 ☐ 1 ☐ 2 X 3 ☐ 4 ☐ 5 ☐ (0 = not at all, 5 = strongly yes)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 6 — Summaries and Overall Assessment

6.1 Diagnosis-oriented evaluation — How successful is the AI mask in lesion delineation compared to GT?

0 ☐ 1 ☐ 2 X 3 ☐ 4 ☐ 5 ☐ (0 = complete failure, 5 = highly successful)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6.2 Prediction-oriented evaluation — How successful is the AI mask in contributing to survival prediction compared to GT?

0 ☐ 1 X 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no added value, 5 = very strong added value)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6.3 Overall confidence in your assessment

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 X 5 ☐ (0 = no confidence, 5 = complete confidence)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6.4 Do you think geometric evaluation metrics (e.g., Dice, IoU, Hausdorff distance) are sufficient to establish clinical trust in AI segmentation?

0 ☐ 1 ☐ 2 ☐ 3 X 4 ☐ 5 ☐ (0 = not sufficient at all, 5 = fully sufficient)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6.5 What additional types of evaluation do you think are necessary for trustworthy AI segmentation (e.g., radiomics stability, survival prediction, physician review, workflow impact, and others), and why do you think those are effective in building trustworthiness?

Reason: \_evaluation of random noises in the segments and fully registered segmentations with the CT as well as an easy-to-use workflow\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 7 — Clinical Adoption & Workflow Impact

7.1 Usability — How easy is it to interpret and use the AI mask compared to GT in clinical workflow?

0 ☐ 1 ☐ 2 x 3 ☐ 4 ☐ 5 ☐ (0 = much harder, 5 = much easier)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.2 Clinical actionability — Would the AI–GT differences change clinical decisions (for example, radiotherapy target, surgery extent, follow-up planning)?

0 ☐ 1 ☐ 2 ☐ 3 x 4 ☐ 5 ☐ (0 = no impact, 5 = strong impact)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.3 Robustness and variability — How consistent and reliable are AI masks across different cases and observers?

0 ☐ 1 ☐ 2 x3 ☐ 4 ☐ 5 ☐ (0 = very inconsistent, 5 = very consistent)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.4 Time-saving potential — How much time could this AI save compared to manual delineation?

0 ☐ 1 ☐ 2 ☐ 3 x 4 ☐ 5 ☐ (0 = no time saved, 5 = saves most effort)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.5 Trust in clinical use — How confident would you be using AI masks for high-stakes tasks (e.g., radiotherapy planning) without physician refinement?

0 ☐1 x 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no trust, 5 = complete trust)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.6 Confidence maps/interpretability (if available) — Do uncertainty or confidence maps improve your trust in AI outputs?

0 ☐ 1 ☐ 2 ☐ 3 x 4 ☐ 5 ☐ (0 = no improvement, 5 = very strong improvement)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.7 Role of AI in clinical practice — In your opinion, AI-based segmentation should:

☐ Replace physicians entirely  
☐ Provide an initial mask for physicians to refine  
x Assist physicians as a support/check tool

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**End**

**Comparative Clinical Assessment of Deep Learning Tumor Masks vs. Ground Truth in Lung Cancer**

Model under review: VNet

**Segmentation Results**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Analysis Mode | Model | Dice Score | IoU Score | Hausdorff Distance |
| Test | **VNet** | 0.81 0.08 | 0.69 0.11 | 11.12 3.73 |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

**Prediction Result:**

Radiomics extracted from Ground Truth used for prediction tasks

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Validation | | | | | | Test | | | | | | | Names | | | |
| Acc | Precision | Recall | F1 | AUC | Specificity | Acc | Precision | Recall | F1 | AUC | Specificity | FSA/AEA | | CA | Model | SL/SSL |
| 0.88 | 0.78 | 0.88 | 0.83 | 0.54 | 1 | 0.87 | 0.76 | 0.87 | 0.81 | 0.71 | 1 | Mutual\_Info\_Gain\_Ratio | | ExtraTree | GT | SSL |
| 0.74 | 0.67 | 0.74 | 0.66 | 0.55 | 0.98 | 0.87 | 0.79 | 0.87 | 0.83 | 0.63 | 0.99 | SL |

Radiomics extracted from Vnet used for prediction tasks

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Validation | | | | | | Test | | | | | | Names | | | |
| Acc | Precision | Recall | F1 | AUC | Specificity | Acc | Precision | Recall | F1 | AUC | Specificity | FSA/AEA | CA | Model | SL/SSL |
| 0.88 | 0.78 | 0.88 | 0.83 | 0.65 | 1 | 0.88 | 0.77 | 0.88 | 0.82 | 0.76 | 1 | LASSO | LGBM | VNet | SSL |
| 0.74 | 0.55 | 0.74 | 0.63 | 0.58 | 0.99 | 0.87 | 0.77 | 0.87 | 0.81 | 0.40 | 0.99 | SL |

**Start**

Section 1 — Overall Impression

1.1 Does the AI mask define a tumor volume that is more clinically meaningful than the Ground Truth (GT) for diagnosis and visual interpretation?

0 ☐ 1 ☐ 2 X3 ☐ 4 ☐ 5 ☐ (0 = not at all, 5 = very strongly yes)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 2 — Spatial Differences

2.1 Are the AI–GT differences concentrated in the peritumoral zone (≈5–15 mm around the tumor)?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 X 5 ☐ (0 = not at all, 5 = strongly concentrated)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

2.2 How clinically meaningful are the regions where the AI mask differs from GT (including over-segmentation or under-segmentation, such as peritumoral expansion, missed necrotic cores, spiculated margins)? Or, how clinically justified are the AI–GT differences (whether due to additional regions or excluded regions)?

0 ☐ 1 ☐ 2 ☐ 3 X 4 ☐ 5 ☐

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 3 — Clinical Interpretation of Differences

3.1 How well does the AI mask preserve internal tumor heterogeneity (for example: solid vs. ground-glass opacity, necrotic vs. viable regions)?

0 ☐ 1 X 2 ☐ 3 ☐ 4 ☐ 5 ☐

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

3.2 Boundary quality: compared to GT, how noisy or fragmented are AI boundaries?

0 ☐ 1 ☐ 2 ☐ 3 X4 ☐ 5 ☐ (0 = much less noisy, 5 = much more noisy)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 4 — Link to Survival Prediction

4.1 Do the AI–GT differences likely capture prognostically relevant tissue that improves survival prediction?

0 X1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

4.2 How would you rate the effect of the AI’s deviation from GT (larger or smaller volume) on survival prediction?

0 ☐ 1 ☐ 2 X 3 ☐ 4 ☐ 5 ☐ (0 = strongly harmful, 5 = strongly beneficial)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 5 — Trust & Preference (Survival Prediction Focus)

5.1 How likely is it that AI-generated masks are more accurate than GT in some cases (for example, subtle lesions missed by physicians, or over-delineation that reduces noise) in survival prediction tasks?

0 ☐ 1 ☐ 2 X3 ☐ 4 ☐ 5 ☐ (0 = very unlikely, 5 = very likely)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

5.2 Would you prefer to use the AI mask (rather than GT) as input for survival prediction?

0 ☐ 1 ☐ 2 X 3 ☐ 4 ☐ 5 ☐ (0 = not at all, 5 = strongly yes)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 6 — Summaries and Overall Assessment

6.1 Diagnosis-oriented evaluation — How successful is the AI mask in lesion delineation compared to GT?

0 ☐ 1 ☐ 2 X 3 ☐ 4 ☐ 5 ☐ (0 = complete failure, 5 = highly successful)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6.2 Prediction-oriented evaluation — How successful is the AI mask in contributing to survival prediction compared to GT?

0 ☐ 1 ☐ 2 ☐ 3 X 4 ☐ 5 ☐ (0 = no added value, 5 = very strong added value)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6.3 Overall confidence in your assessment

0 ☐ 1 ☐ 2 ☐ 3 X 4 ☐ 5 ☐ (0 = no confidence, 5 = complete confidence)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6.4 Do you think geometric evaluation metrics (e.g., Dice, IoU, Hausdorff distance) are sufficient to establish clinical trust in AI segmentation?

0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 X 5 ☐ (0 = not sufficient at all, 5 = fully sufficient)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6.5 What additional types of evaluation do you think are necessary for trustworthy AI segmentation (e.g., radiomics stability, survival prediction, physician review, workflow impact, and others), and why do you think those are effective in building trustworthiness?

Reason: \_evaluation of random noises in the segments and fully registered segmentations with the CT as well as an easy-to-use workflow\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Section 7 — Clinical Adoption & Workflow Impact

7.1 Usability — How easy is it to interpret and use the AI mask compared to GT in clinical workflow?

0 ☐ 1 ☐ 2 x 3 ☐ 4 ☐ 5 ☐ (0 = much harder, 5 = much easier)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.2 Clinical actionability — Would the AI–GT differences change clinical decisions (for example, radiotherapy target, surgery extent, follow-up planning)?

0 ☐ 1 ☐ 2 ☐ 3 x 4 ☐ 5 ☐ (0 = no impact, 5 = strong impact)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.3 Robustness and variability — How consistent and reliable are AI masks across different cases and observers?

0 ☐ 1 ☐ 2 x3 ☐ 4 ☐ 5 ☐ (0 = very inconsistent, 5 = very consistent)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.4 Time-saving potential — How much time could this AI save compared to manual delineation?

0 ☐ 1 ☐ 2 ☐ 3 x 4 ☐ 5 ☐ (0 = no time saved, 5 = saves most effort)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.5 Trust in clinical use — How confident would you be using AI masks for high-stakes tasks (e.g., radiotherapy planning) without physician refinement?

0 ☐1 x 2 ☐ 3 ☐ 4 ☐ 5 ☐ (0 = no trust, 5 = complete trust)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.6 Confidence maps/interpretability (if available) — Do uncertainty or confidence maps improve your trust in AI outputs?

0 ☐ 1 ☐ 2 ☐ 3 x 4 ☐ 5 ☐ (0 = no improvement, 5 = very strong improvement)

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7.7 Role of AI in clinical practice — In your opinion, AI-based segmentation should:

☐ Replace physicians entirely  
☐ Provide an initial mask for physicians to refine  
x Assist physicians as a support/check tool

Reason: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**End**