

Original Research Article

Explainable multimodal deep learning for cancer segmentation and prognostic decision support

Qifeng Wang, Zijie Liu, Chunfeng Wei, Kailin He

Sichuan University Jinjiang College, Meishan, Sichuan, 620860, China

Abstract: This study proposes an explainable multimodal deep-learning framework for cancer segmentation and prognostic decision support, aligned with the research plan of CT/MRI/PET fusion, small-lesion and fuzzy-boundary segmentation, interpretability, generalization, and clinical use. The model combines modality-specific encoders, masked reliability-aware fusion, a boundary-sensitive segmentation branch, and a censored survival branch. Missing examinations are represented by a modality mask. Explanations are organized at spatial, modality, and prognostic-feature levels, with uncertainty-based manual review. Public PET/CT and multimodal MRI cohorts can support development; true CT-MRI-PET fusion requires patient-matched scans and outcomes. The paper defines a reproducible method and validation protocol without claiming unmeasured performance.

Keywords: multimodal deep learning; cancer segmentation; prognosis; explainable artificial intelligence; medical imaging; survival analysis

1. Introduction

Recent 3D segmentation models combine multiscale decoding with transformer representations.^[1] Cancer imaging remains difficult because lesion size, margin clarity, tissue heterogeneity, treatment effects, and acquisition protocols vary substantially. Small or infiltrative lesions are especially difficult to distinguish from surrounding tissue.

CT mainly describes anatomy and density, MRI provides soft-tissue contrast, and FDG-PET reflects glucose metabolism. Their information is complementary. Multicenter PET/CT work has shown that paired anatomical and metabolic imaging can support tumor delineation and outcome modeling.^[2] Multimodal clinical AI must also remain usable when examinations are missing and explanations must be clinically meaningful.^[3]

The method follows the submitted research plan: Preprocessing, multimodal feature extraction, customized loss design, interpretability, generalization, ethical data use, and clinical decision support. Prognosis is modeled as censored time-to-event data, consistent with recent deep-survival research.^[4-5] This is a methodological paper; empirical performance should be reported only after patient-level implementation.

2. Proposed method

2.1. Multimodal preprocessing

CT is resampled to a common spacing and clipped to a task-specific Hounsfield-unit window. MRI undergoes bias-field correction when required and robust normalization. PET is converted to standardized uptake values when metadata permit and percentile-scaled. Patient-matched modalities are registered to a common anatomical space and checked before fusion.

Unrelated public CT, MRI, and PET cohorts are not treated as paired examinations. They may support pretraining, benchmarking, or transfer learning, but patient-level fusion is enabled only when acquisition timing and spatial correspondence meet prespecified criteria. Training combines geometric/intensity

augmentation with modality dropout; a binary mask records available inputs.

2.2. Reliability-aware fusion

Let F_m be the feature map from modality m and M_m in $\{0,1\}$ indicate availability. A reliability module estimates score a_m :

$$w_m = M_m \exp(a_m) / \sum_k M_k \exp(a_k). \quad (1)$$

The fused representation is

$$F_{\text{fusion}} = \sum_m w_m T_m(F_m), \quad (2)$$

where T_m maps each modality into a shared latent space. Cross-modal attention is computed only among available inputs, so missing modalities receive no artificial weight. Reliability weights can reduce the influence of noisy or poorly registered scans, but should be interpreted with controlled modality ablation rather than as causal evidence.^[3,6]

2.3. Boundary-sensitive segmentation

The segmentation branch uses a multiscale decoder with skip connections. To address class imbalance, small lesions, and uncertain margins, the objective is

$$L_{\text{seg}} = \lambda_1 L_{\text{Dice}} + \lambda_2 L_{\text{CE}} + \lambda_3 L_{\text{boundary}}. \quad (3)$$

Dice loss supports sparse foreground learning, cross-entropy preserves voxel-wise discrimination, and the boundary term penalizes contour disagreement. Post-processing remains conservative because removal of small components can improve global overlap while deleting true lesions. Evaluation therefore includes lesion-level sensitivity. Uncertainty may be estimated by test-time augmentation or Monte Carlo dropout; failed registration, severe artifacts, or high boundary uncertainty trigger manual review.

2.4. Prognostic decision support

The prognosis branch pools features from the soft tumor map and a peritumoral region, then combines them with available clinical variables such as age, stage, treatment, and laboratory indicators. Ground-truth contours supervise training but are not deployment inputs.

For right-censored outcomes, a neural proportional-hazards model estimates log-risk $r(x)$:

$$L_{\text{surv}} = -(1/N_e) \sum_{i:\Delta_i=1} [r(x_i) - \log \sum_{j \in R_i} \exp(r(x_j))]. \quad (4)$$

Here Δ_i marks an observed event, R_i is the risk set, and N_e is the number of events. Tied times require explicit handling such as Efron approximation. Cox-based neural models remain widely used for high-dimensional and multimodal survival data.^[4] When survival labels are available, $L = L_{\text{seg}} + \alpha L_{\text{surv}}$. If proportional hazards are violated, a non-proportional or discrete-time survival head should be compared.

The output $r(x)$ is relative risk, not absolute survival probability; fixed-horizon survival requires baseline-hazard estimation and calibration.

2.5. Explainability

Explainability is implemented at three levels: Gradient-based spatial attribution, modality contribution assessed through reliability weights plus ablation, and feature attribution for clinical variables or semantically defined prognostic features. Deep latent variables without stable clinical meaning should not be presented as biomarkers.

Clinical output should combine the tumor mask, uncertainty map, relative risk, contributing modalities, and key prognostic factors. Explanation stability should be checked under perturbation and repeated inference. Recent cancer-XAI reviews note limited clinician involvement; expert review is therefore required.^[6]

3. Validation design

3.1. Datasets and generalization

HECKTOR provides multicenter pre-treatment FDG-PET/CT with tumor annotations and clinical outcomes, supporting paired anatomical-metabolic development of segmentation and prognosis models.^[2] Multimodal MRI brain-tumor data can test transfer to a different imaging environment; recent transformer-based 3D segmentation work provides a technical benchmark.^[1]

These sources must remain separate and cannot form a fictitious CT-MRI-PET cohort. True tri-modal validation requires patient-matched scans, acceptable inter-scan timing, verified registration, and longitudinal outcomes. Data partitions must be patient-independent, including repeated examinations.

Transfer between head-and-neck PET/CT and brain MRI changes disease, anatomy, and modality simultaneously; it is therefore a domain-shift stress test, not isolated proof of cross-cancer generalization. Stronger cross-cancer validation should compare malignancies sharing at least one imaging modality.

3.2. Metrics and ablation

Segmentation should report Dice, 95th-percentile Hausdorff distance, sensitivity, precision, and lesion-level detection sensitivity, preferably stratified by lesion volume. Prognosis should report Harrell's C-index, time-dependent AUC, integrated Brier score, and calibration at prespecified horizons, with patient-level confidence intervals.

Baselines should include unimodal models, early fusion, late fusion, and a competitive recent 3D backbone.^[1] Ablations should remove boundary loss, modality dropout, reliability weighting, cross-modal attention, peritumoral features, and the prognosis branch separately. Missing-modality tests should reproduce clinically plausible combinations.

3.3. Clinical explainability assessment

Spatial explanations can be evaluated using deletion/insertion analysis; modality attribution should be checked against actual modality removal; prognostic explanations should be tested for stability. These outputs describe model evidence, not biological causality.^[3,6]

Radiologists or oncologists can review a prespecified case subset and judge whether explanations reveal segmentation errors, uncertainty is understandable, and risk output is clinically interpretable. Poor registration, severe artifacts, out-of-distribution appearance, or high uncertainty should route cases to manual assessment. The system is decision support rather than autonomous diagnosis.

4. Discussion

The framework addresses a practical limitation of multimodal oncology AI: Complete CT-MRI-PET examinations are uncommon and public datasets rarely provide all modalities with matched outcomes. Modality masks, dropout, and reliability-aware fusion therefore model realistic data availability instead of assuming an ideal three-channel input.

Segmentation quality and prognostic utility are related but not identical. A precise contour can omit peritumoral information associated with recurrence, while survival prediction may use contextual features that do not improve Dice. The prognosis branch therefore uses lesion and peritumoral representations and survival-specific metrics. Recent multimodal cancer-survival literature supports heterogeneous data integration with outcome-specific validation.^[5]

Limitations include registration error, scanner-related shortcut learning, uncertain reliability weights, and survival-model assumptions. Post-hoc explanations do not establish causality. External validation, privacy protection, audit logging, and prospective workflow evaluation are required before deployment, consistent with the research plan's ethical requirements.

5. Conclusion

This paper defines an explainable multimodal deep-learning framework for cancer segmentation and prognostic decision support grounded in CT/MRI/PET fusion, customized segmentation loss, interpretability, generalization, and clinical integration. It explicitly handles missing inputs, emphasizes small-lesion and fuzzy-boundary accuracy, and links segmentation-derived features with censored time-to-event modeling.

The validation strategy separates paired PET/CT development, multimodal MRI transfer testing, and true patient-level CT-MRI-PET validation. No empirical performance is claimed before implementation. Future work should complete patient-level training, external validation, ablation analysis, survival calibration, and clinician-centered explanation assessment.

References

- [1] Tang Y, Yang D, Li W, Roth HR, Landman BA, Xu D, et al. Self-Supervised Pre-Training of Swin Transformers for 3D Medical Image Analysis. *Proc IEEE/CVF Conf Comput Vis Pattern Recognit.* 2022;20730-20740.
- [2] Andrearczyk V, Oreiller V, Boughdad S, et al. Automatic Head and Neck Tumor segmentation and outcome prediction relying on FDG-PET/CT images: Findings from the second edition of the HECKTOR challenge. *Med Image Anal.* 2023;90:102972.
- [3] Pahud de Mortanges A, Luo H, Shu SZ, et al. Orchestrating explainable artificial intelligence for multimodal and longitudinal data in medical imaging. *npj Digit Med.* 2024;7:195.
- [4] Wiegrebe S, Kopper P, Sonabend R, Bischl B, Bender A. Deep learning for survival analysis: a review. *Artif Intell Rev.* 2024;57:65.
- [5] Zhang G, Ma C, Yan C, et al. Multimodal Deep Learning for Cancer Survival Prediction: A Review. *Current Bioinformatics.* 2025;20(4):299-322.
- [6] Mohamed YA, Khoo BE, Mohd Asaari MS, Aziz ME, Ghazali FR. Decoding the black box: Explainable AI for cancer diagnosis, prognosis, and treatment planning-A state-of-the-art systematic review. *Int J Med Inform.* 2025;193:105689.